POLITECNICO DI MILANO

Ph.D. School in Mathematical Models and Methods in Engineering
–XXIV° ciclo –

Polynomial approximation of PDEs with stochastic coefficients

Presented to Dipartimento di Matematica “F. Brioschi”
POLITECNICO DI MILANO

by

Lorenzo TAMELLINI

Graduated in Mathematical Engineering, Politecnico di Milano
matr. 738794

Advisor
Prof. Fabio NOBILE

Tutor:	Prof. Alfio QUARTERONI
Ph.D. coordinator:	Prof. Paolo BISCARI

Milan, March 2012

Abstract

In this thesis we focus on PDEs in which some of the parameters are not known exactly but affected by a certain amount of uncertainty, and hence described in terms of random variables/random fields. This situation is quite common in the engineering practice.

A common goal in this framework is to compute statistical indices, like mean or variance, for some quantities of interest related to the solution of the equation at hand ("uncertainty quantification"). The main challange in this task is represented by the fact that in many applications tens/hundreds of random variables may be necessary to obtain an accurate representation of the solution. The numerical schemes adopted to perform the uncertainty quantification should then be designed to reduce the degradation of their performance whenever the number of parameters increases, a phenomenon known as "curse of dimensionality".

Two methods that seem promising in this sense are the Stochastic Galerkin method and the Stochatic Collocation method. Such methods have therefore recently attracted the interest of the uncertainty quantification community, and have proved to be more effective than sampling methods like Monte Carlo, at least for problems with a moderate number of random parameters. We will compare in detail these methods, and then propose for both suitable generalizations that have shown to be optimal in terms of accuracy per cost for particular problems.

We will also introduce the idea of Generalized Spectral Decomposition for the Stochastic Galerkin method, and explore its application in the context of non scalar equations, focusing on the case of the stationary Navier-Stokes equations. Finally, we will show two applications of the Stochastic Collocation method in the geological and hydraulic engineering.

Keywords: Uncertainty Quantification, Stochastic Galerkin Method, Stochastic Collocation Method, best-M-terms approximation, Generalized Spectral Decomposition

Contents

Introduction

Motivations

The Uncertainty Quantification problem

In the last decades simulation and prediction based on computational models have become widespread tools in science and engineering practice. This process entails the selection of a suitable mathematical model for the system of interest, the tuning of the parameters involved in the model, its discretization and finally its numeric solution. Each of these steps implies some errors (modelling errors, discretization errors, numerical errors, floating point errors); in this thesis we are concerned with the errors that stem from an imperfect knowledge on the system properties.

Indeed, a lack of knowledge on the system of interest occurs in many situations. This may be due to several reasons:

- precision issues in the measurament of some physical quantities like viscosity, permeability, density;
- non-measurability of the system: for example, the permeability of the soil cannot be measured point-wise;
- intrinsic randomness or unpredictability of some quantity, like wind loads, earthquake sources, etc.

We will not deal instead with "model uncertainty", i.e. we will always assume that a suitable mathematical model is available, and the uncertainty is only affecting parameters/coefficients of such model. We will focus on systems that are described by PDEs: in this case the parameters prone to uncertainty are not only the various coefficients (e.g. diffusion, advection, reaction for an elliptic PDE), but also forcing terms, boundary conditions and the shape of the domain.

We will also assume that the uncertainty affecting the parameters can be described through a probabilistic approach. This implies that each uncertain parameter can be modeled as a random variable or random field over a complete probability space, and that the probability distribution of such random objects is known, either from experimental measures or from human expertise: this of course is a not-so-small assumption in practice and proper statistical inference tools should be used to characterize the randomness in the system starting from available measuraments. Again, this aspect is not addressed here.

As soon as some of the parameters of the equation depend on a random event, so does the solution: each realization of the random parameters will correspond to a different solution, through the evaluation of a solver function, that in the case of PDEs usually requires assemblying and solving the linear system coming e.g. from a finite difference or finite element discretization.

In this context, the goal is usually to compute statistics of the solution, like mean, variance, probability of exceeding a threshold value ("failure probability") etc; often one could also be interested in restricting such statistical analysis to functionals of the solution (hereafter called "output quantities" or "quantities of interest"). This kind of analysis is usually referred to as "Uncertainty Quantification " or "Forward Uncertainty Propagation" from the inputs to the outputs of the model.

The sampling approach to Uncertainty Quantification

Monte Carlo sampling is the most natural approach to solve the Uncertainty Quantification problem, see e.g. [16, 87] . This simply requires generating a set of independent realizations from the parameters space, and solving the equation for each sample in the set; the statistical indices are then approximated by sample averages over the obtained set of solutions. Such method has a straightforward implementation, since one can readily exploit all pre-existing deterministic codes, to be used as "black boxes", and moreover is fully parallelizable; see e.g. [51, 57, 70] for a review on random points generators.

The convergence of the Monte Carlo method can be proved using classical results of probability theory, that is the *Law of Large Numbers* and the *Central Limit Theorem*, see e.g. [63, 64, 89]. In practice very few assumptions are required on the structure of the equation itself for the Monte Carlo method to converge.

The main drawback of the Monte Carlo method is the extremely slow rate of convergence, which, due to the *Central Limit Theorem*, is only proportional to $\sigma/\sqrt{M}$, where M denotes the number of samples and σ is the standard deviation of the considered quantity of interest (considered bounded). This can result in massive computational cost when the cost of a single evaluation of the solver function requires complex operations like solving a linear system. Such rate is however independent of the number of random parameters considered; this is a desirable feature since, as we will see later, the number of random parameters involved can be quite large, and many methods suffer from a degradation of their performance as this numbers increases, a phenomenon known as "Curse of Dimensionality".

A number of methods have been proposed to improve the convergence of Monte Carlo method, either in the context of random sampling or considering deterministically chosen points. In the former case, an improvement in the performances with respect to the standard Monte Carlo method is achieved either improving the random sampling efficiency (Stratified Sampling, Latin Hypercube Sampling), or through variance reduction techniques (Antithetic Variates, Control Variates, Importance Sampling), see e.g. [16, 88]. All of these methods are quite effective in improving the constant σ in the convergence estimate of the Monte Carlo method, but only the Latin Hypercube Sampling improves the convergence rate 1/2 up to 1, see e.g. [65, 98].

On the other hand, in a deterministic framework Quasi-Monte Carlo methods have been proposed, see [16, 52, 70, 94]. In these methods the sample indeed is not randomly generated, but the points are deterministically chosen (Halton points, Sobol' points) so to maximize some suitable measure of efficiency in the coverage of the parameters space (the so-called "low-discrepancy" property). These methods can achieve convergence rate up to 1 (up to logarithmic terms), depending on the smoothness of the map from the parameters to the quantity of interest ("response surface"). To this end, we also mention the recent developments in [26].

Polynomial approximation of the response surface

If the response surface (i.e. the input-output map) is regular one could exploit this property in order to reduce the computational burden, computing for instance a polynomial approximation of it. Such approximation is also known in literature as "surrogate model" and has been introduced mainly for optimization purposes, see e.g. [55]. As we will see, once the surrogate model has been computed, statistics like mean or variance can be approximated with simple post-process with almost no computational cost, using suitable quadrature formulae.

In this thesis we will mainly consider systems governed by elliptic PDEs, for which it is possible to prove analyticity of the response surface (see Chapters 2 and 3), therefore looking for its polynomial approximation is sound. Yet, the construction of the approximated response surface will be in general a difficult task, since the quantity of interest may depend on a high number of random parameters.

We will focus on two types of polynomial approximations: interpolant schemes, called Stochastic Collocation, and projection ones, called Stochastic Galerkin. Both interpolation and projection are well-established and efficient techniques exist for the approximation of real-valued functions of one variable: their efficient extension to the approximation of high dimensional response surfaces has been a central issue for the Uncertainty Quantification community in the last decades, and will be the focus of this thesis.

Outline

This Thesis is organized as follows:

Chapter 1: exposes concisely the main results of this work, putting in evidence its main thread and leaving details to the subsequent Chapters.

Chapter 2: compares the Stochastic Galerkin and Collocation methods and their respective performances. We also introduce and explore numerically a first idea for anisotropic approximation schemes.

Chapter 3: investigates in detail the structure of the response surface for an ellliptic PDE with random diffusion coefficient, and exploits this theoretical understanding to derive particular versions of the Stochastic Galerkin and Collocation method that are quasi-optimal with respect to a classical accuracy-cost ratio criterion.

Chapter 4: applies the optimal collocation technique derived in Chapter 3 on a groundwater flow problem.

Chapter 5: recasts the Stochastic Galerkin approximation procedure into a generalized eigenvalue problem, and uses this fact to build a sequence of surrogate models that converges to the complete Galerkin approximation. Since computing each element of the sequence is much cheaper than solving a full Galerkin problem, such procedure allows to obtain decent approximations of the Galerkin solution with remarkable savings in terms of computational cost with respect to the standard Galerkin method. We apply this technique to a stationary Navier-Stokes problem, with uncertainty on Reynolds number and forcing field.

Chapter 6: develops tools for the computation of some sensitivity indices, that will be used to perform an Uncertainty Quantification analysis on a geochemical compaction problem.

Most of the chapters are based on works already published/accepted for publication, or ready for submission, and they are therefore “self-contained”. For these chapters, we have decided to keep the structure of the corresponding paper with only minor changes, even if this implies some repetitions:

Chapter 2: J. Bäck, F. Nobile, L. Tamellini, R. Tempone *Stochastic Spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison* J.S. Hesthaven and E.M. Ronquist, editors, *Spectral and High Order Methods for Partial Differential Equations*, volume 76 of *Lecture Notes in Computational Science and Engineering*, pages 43-62. Springer, 2011. Selected papers from the ICOSAHOM 09 conference, June 22-26, Trondheim, Norway.

Chapter 3: J. Beck, F. Nobile, L. Tamellini, R. Tempone *On the optimal polynomial approximation of stochastic PDEs by Galerkin and Collocation methods* To appear on *Mathematical Models and Methods in Applied Sciences.* Also available as MOX report 23/2011 - Department of Mathematics, Politecnico di Milano.

A shorter version can also be found on J. Beck, F. Nobile, L. Tamellini, R. Tempone, *Implementation of optimal Galerkin and Collocation approximations of PDEs with random coefficients*, in ESAIM: Proceedings, 33 (2011). Proceedings of CANUM 2010, Carcans-Maubuisson, France, May 31-June 4, 2010.

Chapter 5: L. Tamellini, O.P. Le Maitre, A. Nouy *Generalized stochastic spectral decomposition for the Steady Navier–Stokes equations*, in preparation.

Chapter 6 is also based on a paper in preparation, *A numerical model for the geological compactation of sedimentary basins with sensitivity analysis*, by L. Formaggia, A. Guadagnini, I. Imperiali, G. Porta, M. Riva, A. Scotti, L. Tamellini. However, since the focus of the paper is more on the geological application, in this case we have rewritten most of the content of the paper, focusing mostly on the Uncertainty Quantification aspects.

Most of this work has been carried out at MOX laboratory, Department of Mathemtatics, Politecnico di Milano.

Chapters 2,3 have been also developed during multiple stays visiting prof. Raul Tempone and his team, both at ICES laboratory, University of Texas at Austin, USA, and Applied Mathematics and Computational Science Department, King Abdullah University for Science and Technology (KAUST) at Thuwal, Saudi Arabia.

Chapter 5 was mainly developed throughout a visit to dr. Olivier Le Maitre at LIMSI-CNRS laboratory, Université Paris-Sud 11 at Orsay, France, during winter 2010-2011.

Chapter 6 stems from a collaboration between MOX laboratory, Hydraulic and Environmental Department (DIIAR) at Politecnico di Milano and ENI.

All the computational results shown in this work have been performed in MATLAB® language, except from Chapter 5, in which the code has been written in C++ and MATLAB® has been used only for post-processing.

Acknowledgements

This thesis is the result of the fruitful collaboration among many people, who deserve grateful acknowledgement.

My deepest thanks goes to my advisor, Fabio Nobile, for the huge support in every aspect of this three-years-long work, from the insipiring discussions at the blackboard to the technical guidance and suggestions in the analysis of theorems and numerical results, to the proof-reading of every single page and slide I have written, with 24/7 availability.

A huge thanks goes also to Raul Tempone: in addition to the warm ospitality in the many places I have visited him and the contagious enthusiasm in facing any mathematical challenge, I always have greatly benefitted from his help in discussing and dissecting any problem we had to solve and from his many suggestions.

I am also greatly thankful to Olivier Le Maitre for the nice opportunity of visiting him in Paris. I have really appreciated the great support in understanding the problems we were tackling and in the writing and running of the code, the overall help and the patience.

At Politecnico di Milano, I would like to thank Luca Formaggia, Alberto Guadagnini, Monica Riva and all the people at MOX lab and DIIAR department involved in the geological compaction problem.

A very special thanks goes to Joakim Beck, for having been not only a precious help in discussing, coding, debugging, and running codes, but also a very good friend.

This thesis would not have been possible without the amazing and constant support I got from my beloved ones: my parents, my brother, Nadia and Ferruccio. The grand finale is however for Rossana, who is my love, my strength, my hope and my change of perspective. This thesis is dedicated to her, and I cannot think of better ways to write the last sentence of this work.

Chapter 1

Thesis overview

This Chapter shortly highlights the main results contained in this Thesis, putting in evidence the main thread of the work and pointing to the subsequent Chapters for the full discussion. All the necessary background and notation is confined to Section 1.1.

1.1 Problem setting

Let us consider the linear elliptic equation

$$\begin{cases} -\operatorname{div}(a\nabla u) = f & \mathbf{x} \in D = (0,1)^2, \\ u = 0 & \mathbf{x} \in \partial D, \end{cases} \tag{1.1}$$

that describes physical phenomena like heat transfer or Darcy's flows in porous media, see e.g. [8, 50]. A straightforward application of the Lax–Milgram lemma allows to conclude that equation (1.1) admits a unique solution u in $H^1_0(D)$, the space of square integrable functions in D with square integrable distributional derivatives and zero trace on the boundary, provided f is in the dual space of $H^1_0(D)$, which we denote by $H^{-1}(D)$.

As pointed out in the introduction, we are concerned with situations in which the parameters of (1.1) are affected by uncertainty. For easiness of presentation we consider the case where the coefficient a in (1.1) is the only source of uncertainty in the model, but the theory we present extends immediately to the more general case where also f and the boundary conditions are uncertain. The shape of D can also be considered as uncertain, see e.g. [18, 46, 47, 77, 112].

We describe the uncertainty on a with a probabilistic approach. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space: Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events and $P : \mathcal{F} \to [0,1]$ is a probability measure. We recall that a real-valued random variable on $(\Omega, \mathcal{F}, P)$ is a function $X = X(\omega) : \Omega \to \mathbb{R}$ that assigns a numerical value to each outcome $\omega \in \Omega$. We denote with $\mathbb{E}\left[X^k\right]$ the k-th order moment of the random variable,

$$\mathbb{E}\left[X^k\right] = \int_\Omega X^k(\omega) dP(\omega).$$

In particular $\mathbb{E}[X]$ denotes the mean, or expected value of X, while the variance of X is defined as $\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. Let moreover $L^2_P(\Omega)$ be the space of random variables with bounded second moments,

$$L^2_P(\Omega) = \left\{ X(\omega) : \mathbb{E}\left[X^2\right] = \int_\Omega X^2(\omega) dP(\omega) < \infty \right\}.$$

We further denote by $\mathbb{C}\mathrm{ov}\left[XY\right]$ the covariance between two random variables X and Y

$$\mathbb{C}\mathrm{ov}\left[XY\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right]. \tag{1.2}$$

Next we state some assumptions on the diffusion coefficient a. These are indeed quite strong assumptions that will be weakened later on, when dealing with the so-called "lognormal" random fields, see e.g. Chapter 4.

Assumption 1.1. *The diffusion coefficient $a = a(\mathbf{x}, \omega)$ is a random field on $(\Omega, \mathcal{F}, P)$ taking values in $L^\infty(D)$, i.e. a function from $\overline{D} \times \Omega$ $\mathbb{R}$ such that*

1. *$a(\cdot, \omega)$ is a strictly positive and bounded function over D for each random event $\omega \in \Omega$, i.e. there exist two positive costants $\infty > a_{max} > a_{min} > 0$ such that*

$$P(a_{min} \leq a \leq a_{max}) = 1,$$

2. *$a(\mathbf{x}, \cdot)$ is a real-valued random variable for each point in $\overline{D}$.*
3. *for $\mathbf{p}, \mathbf{q} \in D$, the covariance function*

$$C_a(\mathbf{p}, \mathbf{q}) = \mathbb{C}\mathrm{ov}\left[a(\mathbf{p}, \cdot)a(\mathbf{q}, \cdot)\right] \tag{1.3}$$

depends only on the distance $\|\mathbf{p} - \mathbf{q}\|$ ("weak stationarity" property).

Since to every realization of a corresponds a different solution $u \in H^1_0(D)$, u is in turn a random field on $(\Omega, \mathcal{F}, P)$, taking values in $H^1_0(D)$. Equation (1.1) is therefore understood to hold in a P-almost everywhere sense:

Strong Formulation. *find a $H^1_0(D)$-valued random field, $u : \overline{D} \times \Omega \to \mathbb{R}$, such that P-almost everywhere in Ω, or in other words almost surely (a.s.), the following equation holds:*

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \omega)\nabla u(\mathbf{x}, \omega)) = f(\mathbf{x}) & \mathbf{x} \in D, \\ u(\mathbf{x}, \omega) = 0 & \mathbf{x} \in \partial D, \end{cases} \tag{1.4}$$

where the operators div *and ∇ imply differentiation with respect to the physical coordinate only.*

The Proposition stated next is an immediate consequence of Assumption 1.1 and Lax–Milgram Lemma.

Proposition 1.1. *For any $f \in H^{-1}(D)$, the strong formulation 1.4 admits a unique solution $u(\cdot, \omega) \in H^1_0(D)$ for almost every $\omega \in \Omega$, with*

$$\|u(\cdot, \omega)\|_{H^1(D)} \leq \frac{1}{a_{min}} \|f\|_{H^{-1}(D)},$$

Moreover, for any $\mathbf{x} \in D$, $u(\mathbf{x}, \cdot) \in L^q_P(\Omega)$, for all $q \in \mathbb{N}$, $1 \leq q \leq \infty$.

As anticipated in the Introduction, the aim of an Uncertainty Quantification analysis is to compute statistical indices for u, like $\mathbb{E}\left[u\right]$ or $\mathbb{V}\mathrm{ar}\left[u\right]$, or a failure probability $P(u > u_0)$. Often one could also be interested in performing the same analysis for a linear or non linear functional of u, $\psi(u) : H^1_0(D) \to \mathbb{R}$, that represents the quantity of interest for the Uncertainty Quantification analysis. Note that, under the assumption of linearity or Lipschitz continuity of ψ, Proposition 1.1 implies that the moments of ψ are bounded, see e.g. [71].

1.1.a The Finite Dimensional Noise Assumption

Suppose for a moment that the following additional assumption on the diffusion coefficient holds:

Assumption 1.2 ("Finite Dimensional Noise Assumption")**.** *$a(\mathbf{x}, \omega)$ can be represented by a vector of N real-valued random variables $\mathbf{y} = (y_1, \ldots, y_N)^T$,*

$$a(\mathbf{x}, \omega) = a(\mathbf{x}, y_1(\omega), y_2(\omega), \ldots, y_N(\omega)\,).$$

More general situations will be addressed in next subsection. Without loss of generality, we can also consider each y_i to have zero mean and unit variance and, for easiness of presentation, we will also consider them identically distributed. Let $\Gamma_i \subseteq \mathbb{R}$ be the support of y_i, $\Gamma = \Gamma_1 \times \Gamma_2 \times \ldots \times \Gamma_N$ the support of $\mathbf{y}$ and $\rho(\mathbf{y})$ the joint probability density function of $\mathbf{y}$, such that

$$P(y_1 < t_1, y_2 < t_2, \ldots, y_N < t_N) = \int_{-\infty}^{t_1} \int_{-\infty}^{t_2} \cdots \int_{-\infty}^{t_N} \rho(\mathbf{y}) d\mathbf{y}.$$

We can therefore replace the abstract probability space $(\Omega, \mathcal{F}, P)$ with $(\Gamma, \mathcal{B}(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, where $\mathcal{B}(\Gamma)$ denotes the Borel σ-algebra, and the space $L^2_P(\Omega)$ with $L^2_\rho(\Gamma)$, defined as

$$L^2_\rho(\Gamma) = \left\{ f : \Gamma \to \mathbb{R} \text{ s.t. } \int_\Gamma f^2(\mathbf{y})\rho(\mathbf{y})d\mathbf{y} < \infty \right\}.$$

Furthermore, once a has a representation in terms of N random variables u can also be expressed in terms of the same random variables, $u = u(\mathbf{x}, \mathbf{y})$ with $u(\mathbf{x}, \cdot) \in L^2_\rho(\Gamma)$. Therefore the strong formulation (1.4) now reads:

Strong Formulation (finite dimensional)**.** *find $u : \overline{D} \times \Gamma \to \mathbb{R}$ such that $\rho(\mathbf{y})d\mathbf{y}$-almost everywhere in Γ it holds:*

$$\begin{cases} -\operatorname{div}(a(\mathbf{x}, \mathbf{y})\nabla u(\mathbf{x}, \mathbf{y})) = f(\mathbf{x}) & \mathbf{x} \in D, \\ u(\mathbf{x}, \mathbf{y}) = 0 & \mathbf{x} \in \partial D. \end{cases} \tag{1.5}$$

With a view to the polynomial approximation of the map $\mathbf{y} \to u(\mathbf{x}, \mathbf{y})$, it is also useful to introduce a weak formulation of (1.5). Proposition 1.1 suggests that an appropriate space for u is the tensor space $H^1_0(D) \otimes L^2_\rho(\Gamma)$, thus assuming a separation between the physical and stochastic variables,

$$u = \sum_{i \in \mathbb{N}} u_i(\mathbf{x})\varphi_i(\mathbf{y}) \quad u_i \in H^1_0(D),\, \varphi_i \in L^2_\rho(\Gamma), \tag{1.6}$$

so that the weak formulation of (1.5) can be stated as

Weak Formulation. *find $u \in H^1_0(D) \otimes L^2_\rho(\Gamma)$ such that $\forall v \in H^1_0(D) \otimes L^2_\rho(\Gamma)$*

$$\int_\Gamma \int_D a(\mathbf{x}, \mathbf{y})\nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \tag{1.7}$$

In the following we will often exploit the twofold interpretation of u, either as a map from $\mathbf{y} \in \Gamma$ to $u(\mathbf{y}) \in H^1_0(D)$ or as a function in the tensor space $H^1_0(D) \otimes L^2_\rho(\Gamma)$, using in each case the most convenient one.

It is immediate to introduce a discretization in space of the Weak Formulation (1.7). In this thesis we have considered a finite element discretization, but finite difference or finite Volume could be employed as well. Denoting with $\mathcal{T}_h$ a triangulation of the physical domain D and with $V_h(D) \subset H^1_0(D)$ a finite element space of piecewise continuous polynomials on $\mathcal{T}_h$, whose dimension is N_h, we can write a weak formulation for the semi-discrete problem in space as

Weak Formulation (semidiscrete). *find* $u_h \in V_h(D) \otimes L^2_\rho(\Gamma)$ *such that* $\forall v \in V_h(D) \otimes L^2_\rho(\Gamma)$

$$\int_\Gamma \int_D a(\mathbf{x},\mathbf{y}) \nabla u_h(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{x},\mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x},\mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \tag{1.8}$$

At this point it is very convenient to further assume independence of the random variables y_i introduced with the Finite Dimensional Noise Assumption 1.2:

Assumption 1.3. *The random variables* y_i *are mutually statistically independent.*

Under such Assumption indeed the joint probability density function $\rho(\mathbf{y})$ factorizes as $\rho(\mathbf{y}) = \prod_{i=1}^N \rho_i(y_i)$. Therefore $L^2_\rho(\Gamma) = \bigotimes_{i=1}^N L^2_{\rho_i}(\Gamma_i)$, and a basis for $L^2_\rho(\Gamma)$ can be built taking products of basis functions for $L^2_{\rho_i}(\Gamma_i)$. We remark however that the assumption of independence of y_i is not essential, and it is possible to work within a tensor structure also in the case of non independent y_i by introducing an auxiliary density $\hat{\rho}$ that factorizes, as proposed in [4]. The price to pay in the convergence estimates is then a constant factor proportional to $\|\rho/\hat{\rho}\|_{L^\infty(\Gamma)}$.

1.1.b Expansions of a random field

In some cases the assumptions of finite dimensional noise and independence of the random parameters y_i are natural for the problem at hand; think e.g. to a composite medium where each subdomain has its own random diffusion coefficient. However, in many situations a finite dimensional structure is not readily available, and one has to resort to decomposition and truncation techniques to obtain an approximate representation of a with a finite number of random variables. Such number has to be large enough to take into account a sufficient amount of the total variability of the field, defined as $\int_D \mathbb{V}\mathrm{ar}\left[a(\mathbf{x},\cdot)\right] d\mathbf{x}$ [1]. We remark that in many applications of interest, this number may be of order of tens/hundreds, see e.g. Chapter 4, hence the need for careful techniques in building the polynomial approximation of the solution u of (1.8).

The most common decomposition of a random field is perhaps the Karhunen-Loève expansion, see [42, 62, 63, 64], which is the continuous anologous of the Principal Component Decomposition (see e.g. [54]). In the Karhunen-Loève expansion the random field is expanded in terms of uncorrelated[2] random variables y_i, $L^2(D)-$orthonormal deterministic functions $v_i(\mathbf{x})$ and decreasing coefficients λ_i,

$$a(\mathbf{x},\omega) \approx a_N(\mathbf{x},\omega) = \mathbb{E}\left[a(\mathbf{x},\cdot)\right] + \sum_{i=1}^N \sqrt{\lambda_i} v_i(\mathbf{x}) y_i(\omega), \tag{1.9}$$

see Section 4.3 for a full discussion on such expansion. As an alternative to the Karhunen-Loève expansion, it is possible to compute a Fourier-based decompostion of a, which uses trigonometric polynomials as basis functions in the physical space, thus highlighting the contribution of each frequency to the total field a. Both expansion have similar convergence properties, a smoother covariance function (1.3) resulting in a faster convergence.

The KL and Fourier expansions result in a linear dependence of the random field a on the random variables y_i. In some applications however it is useful to model the logarithm of a as a random field rather than a itself (usually when modeling phenomena with large variations), so that the final expression for a depends exponentially on y_i:

$$a_N(\mathbf{x},\omega) = a_0(\mathbf{x}) + \exp\left(\sum_{i=1}^N c_i v_i(\mathbf{x}) y_i(\omega)\right). \tag{1.10}$$

[1] A way of working without formally dropping any term in the expansion of a will be addressed in Chapter 3.

[2] Observe that if the random field is Gaussian then y_i are Gaussian too and therefore independent.

Such expansion guarantees almost sure positivity of $a(\mathbf{x}, \omega)$, a property useful when describing diffusion coefficients or other physical parameters that need to be positive. Such logarithm expansion is usually considered together with the assumption that y_i are Gaussian random variables, thus obtaining the so-called "lognormal model". Note that in this case a is not uniformely bounded with respect to ω, nor coercive in the case $a_0 = 0$, so that Assumption 1.1 is not satisfied. However, it is still possible to prove the well-posedness of the strong and weak formulations (1.4)-(1.8), as shown in [19, 43].

1.1.c Methods for polynomial approximation of u

We are now in position to define the polynomial approximation (surrogate model) for the map $\mathbf{y} \in \Gamma \to u_h(\mathbf{y}) \in V_h(D)$ that solves (1.8). We introduce a polynomial subspace of $L^2_\rho(\Gamma)$, which we denote by $\mathbb{P}(\Gamma)$, and look for a full discrete solution $u_{h,w} \in V_h(D) \otimes \mathbb{P}(\Gamma)$ solving[3]

Weak Formulation (Fully discrete). *find* $u_{h,w} \in V_h(D) \otimes \mathbb{P}(\Gamma)$ *such that* $\forall v \in V_h(D) \otimes \mathbb{P}(\Gamma)$

$$\int_\Gamma \int_D a_N(\mathbf{x}, \mathbf{y}) \nabla u_{h,w}(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}, \tag{1.11}$$

with the understanding that the polynomial space $\mathbb{P}(\Gamma)$ should be designed to retain good approximating properties while keeping the number of degrees of freedom as low as possible. In this sense, using the classical Tensor Product polynomials space, that contains all the N-variate polynomials with maximum degree lower than a given $w \in \mathbb{N}$, is not a good choice, since its dimension grows exponentially fast with the number of random variables N, $\dim \mathbb{P}(\Gamma) = (1 + w)^N$. A valid alternative choice that has been widely used in literature (see e.g. [42, 97, 111]) is to use the Total Degree polynomial space, that includes the polynomials whose sum of degrees in each variable is lower than or equal to w: such space contains indeed only $\binom{N+w}{N}$ polynomials, which is much lower than $(1 + w)^N$, yet retaining good approximation properties, as will be shown in the next sections. A number of polynomial spaces is listed and analyzed in Chapter 2. One could also introduce anisotropy techniques, with the aim to enrich the polynomial space only in those direction of the stochastic space which are seen to contribute most to the total variability of the solution.

In practice, the fully discrete solution $u_{h,w}$ will be computed either with a projection on $V_h(D) \otimes \mathbb{P}(\Gamma)$, resulting in the Galerkin method, or as an intepolant, resulting in the Collocation Method. We will now briefly recall the construction of such methods, and then move to the results of this thesis.

1.1.d Stochastic Galerkin method

As anticipated, the Stochastic Galerkin method (see e.g. [3, 42, 66, 101, 60]) aims at computing the modal coefficients of $u_{h,w}$ in $V_h(D) \otimes \mathbb{P}(\Gamma)$, i.e. the coefficients of the expansion $u_{h,w}(\mathbf{x}, \mathbf{y}) = \sum_{p=1}^M u_p(\mathbf{x}) \varphi_p(\mathbf{y})$ for a suitable basis $\{\varphi_p\}_{p=1,\dots,M}$ of $\mathbb{P}(\Gamma)$. It is then convenient to endow $\mathbb{P}(\Gamma)$ with a $\rho(\mathbf{y})d\mathbf{y}$-orthonormal basis, i.e. with a sequence of polynomials $\{\mathcal{L}_p\}_{p\in\mathbb{N}}$ such that $\int_\Gamma \mathcal{L}_p \mathcal{L}_q \rho(\mathbf{y})\, d\mathbf{y} = 1$ if $p = q$ and 0 otherwise. To this end we take advantage of the tensor structure of $L^2_\rho(\Gamma)$ deriving from Assumption 1.3, and build the elements of such basis as products

[3] Note that since the diffusion coefficient a_N may have been obtained by a truncation of a random field, the solution of the problem actually also depends on the truncation parameter N, and hence should be denoted by $u_{N,h,w}$. However, since the focus of this thesis is not on the convergence with respect to N, we omit the corresponding subscript, for the sake of simplicity in the notation.

of $\rho_i(y_i)dy_i$-orthonormal polynomials, which we denote as $\{L_p\}_{p\in\mathbb{N}}$:

$$\mathcal{L}_{\mathbf{p}}(\mathbf{y}) = \prod_{n=1}^{N} L_{p_n}(y_i) \quad \mathbf{p} \in \mathbb{N}^N.$$

Families of $\rho_i(y_i)dy_i$-orthonormal polynomials exist for many probability distribution: we recall Legendre polynomials for uniform measures and Hermite polynomials for Gaussian measures (see [111] for the general Askey scheme), for which explicit formulae and computing algorithms are available, see e.g. [39]. See also [31] for examples of probability measures that do not admit such an orthonormal basis.

To allow for general polynomial spaces we introduce a sequence of increasing index sets $\Lambda(w)$ such that

$$\Lambda(0) = \{(0,\dots,0)\}, \quad \Lambda(w) \subseteq \Lambda(w+1) \subset \mathbb{N}^N \text{ for } w \geq 0, \quad \mathbb{N}^N = \bigcup_{w\in\mathbb{N}} \Lambda(w), \tag{1.12}$$

each with cardinality M_w, and consider the corresponding polynomial space

$$\mathbb{P}_{\Lambda(w)}(\Gamma) = span\,\{\mathcal{L}_{\mathbf{p}}(\mathbf{y}),\ \mathbf{p} \in \Lambda(w)\} \tag{1.13}$$

for the approximation of $u_{h,w}$ with the Galerkin method. In other words, the Galerkin method will compute the coefficients $u_{\mathbf{p}} \in V_h(D)$ of the expansion

$$u_{h,w}(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{p}\in\Lambda(w)} u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y}). \tag{1.14}$$

Such expansion is usually known as generalized Polynomial Chaos Expansion (gPCE). Having the gPCE expansion of $u_{h,w}$ (1.14) allows one to compute easily the mean and variance of $u_{h,w}$ as

$$\mathbb{E}\left[u_{h,w}(\mathbf{x},\cdot)\right] = u_{\mathbf{0}}(\mathbf{x}), \qquad \mathbb{V}\mathrm{ar}\left[u_{h,w}(\mathbf{x},\cdot)\right] = \sum_{\mathbf{p}\in\Lambda(w)} u_{\mathbf{p}}^2(\mathbf{x}) - u_0^2.$$

The final step of the Galerkin method is to further consider the FEM approximation of the coefficients $u_{\mathbf{p}}(\mathbf{x})$ in (1.14) and insert it in the fully discrete weak formulation (1.11). This will result in a set of $N_h \times M_w$ linear equations that couple all modes $u_i(\mathbf{x})$, due to the presence in (1.11) of terms like $\int_{\Gamma_i} y_i L_{p_i}(y_i) L_{q_i}(y_i)\rho_i(y_i)dy_i$; see Chapter 2 for more details on the discrete problem.

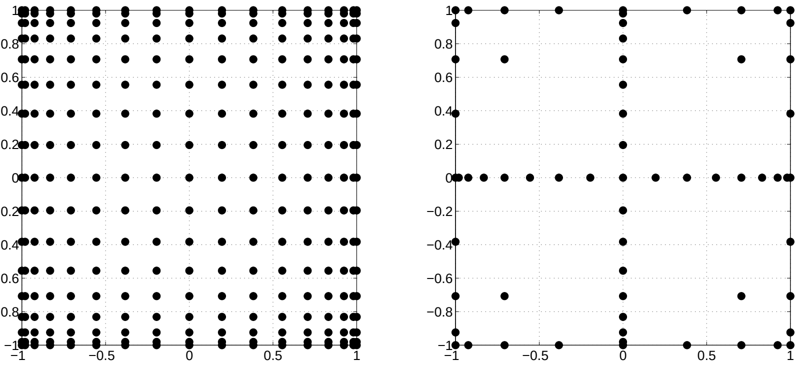
1.1.e Stochastic Collocation method

As an alternative to the Stochastic Galerkin modal approach, the Stochastic Collocation method ([4, 38, 110]) consists in collocating problem (1.5) in a set of Q points $\{\mathbf{y}_j \in \Gamma\}$, i.e. computing the corresponding solutions $u(\cdot,\mathbf{y}_j)$ and building a global polynomial approximation $u_{h,w}$, not necessarily interpolatory, upon those evaluations: $u_{h,w}(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{Q} u(\mathbf{x},\mathbf{y}_j)\mathscr{P}_j(\mathbf{y})$ for suitable multivariate polynomials $\{\mathscr{P}_j\}_{j=1}^{Q}$.

Building the set of evaluation points $\{\mathbf{y}_j\}$ as a cartesian product of monodimensional grids becomes quickly unfeasible, since the computational cost grows exponentially fast with the number of stochastic dimensions needed (see figure 1.1(a)). We consider instead the so-called *sparse grid* procedure (see figure 1.1(b)), originally introduced by Smolyak in [95] for high dimensional quadrature purposes; see also [7, 15] for polynomial interpolation.

For each direction y_n we introduce a sequence of one dimensional polynomial Lagrangian interpolant operators of increasing order:

$$\mathcal{U}_n^{m(i)} : C^0(\Gamma_n) \to \mathbb{P}_{m(i)-1}(\Gamma_n).$$



(a) Tensor grid

(b) Sparse grid

Figure 1.1: Tensor grid 1.1(a) vs. Sparse grid 1.1(b).

Here $i \geq 1$ denotes the level of approximation and $m(i)$ the number of collocation points used to build the interpolation at level i. We require the function m to satisfy the following assumptions:

$$m(0) = 0, \quad m(1) = 1, \quad m(i) < m(i+1), i \geq 1.$$

In addition, let $\mathcal{U}_n^0[q] = 0, \forall q \in C^0(\Gamma_n)$. Next we introduce the difference operators $\Delta_n^{m(i)} = \mathcal{U}_n^{m(i)} - \mathcal{U}_n^{m(i-1)}$, an integer value $w \geq 0$, multi-indices $\mathbf{i} \in \mathbb{N}_+^N$ and a sequence of index sets $\mathcal{I}(w)$ such that $\mathcal{I}(w) \subset \mathcal{I}(w+1)$ and $\mathcal{I}(0) = \{(1, 1, \ldots, 1)\}$. We define the sparse grid approximation of $u_h : \Gamma \to V_h(D)$ at level w as

$$u_{h,w}(\mathbf{y}) = \mathcal{S}_{\mathcal{I}(w)}^m[u_h](\mathbf{y}) = \sum_{\mathbf{i} \in \mathcal{I}(w)} \bigotimes_{n=1}^N \Delta_n^{m(i_n)}[u_h](\mathbf{y})\,. \tag{1.15}$$

As pointed out in [41], it is desiderable that the sum (1.15) has some telescopic properties. To ensure this we have to impose some additional constraints on $\mathcal{I}$. Following [41] we say that a set $\mathcal{I}$ is *admissible* if $\forall\, \mathbf{i} \in \mathcal{I}$

$$\mathbf{i} - \mathbf{e}_j \in \mathcal{I} \text{ for } 1 \leq j \leq N, i_j > 1. \tag{1.16}$$

We refer to this property as *admissibility condition*, or *ADM* in short. Given a set $\mathcal{I}$ we will denote by $\mathcal{I}^{ADM}$ the smallest set such that $\mathcal{I} \subset \mathcal{I}^{ADM}$ and $\mathcal{I}^{ADM}$ is admissible.

The set of all evaluation points needed by (1.15) is called *sparse grid*, and has cardinality Q. Note that (1.15) is indeed equivalent to a linear combinations of tensor grids interpolations, each of which contains "few" interpolation points:

$$u_{h,w}(\mathbf{y}) = \sum_{\mathbf{i} \in \mathcal{I}(w)^{ADM}} c_i \bigotimes_{n=1}^N \mathcal{U}_n^{m(i_n)}[u_h](\mathbf{y}), \quad c_i = \sum_{\substack{\mathbf{j} = \{0,1\}^N \\ \mathbf{i}+\mathbf{j} \in \mathcal{I}(w)^{ADM}}} (-1)^{|\mathbf{j}|}. \tag{1.17}$$

Observe that many coefficients c_i in (1.17) are zero. Once the Stochastic Collocation approximation has been obtained, the computation of moments of $u(\mathbf{x}, \cdot)$ simply requires the application of a quadrature rule that naturally derives from equation (1.17),

$$\mathbb{E}\left[u_{h,w}(\mathbf{x}, \cdot)\right] = \sum_{j=1}^Q u_h(\mathbf{x}, \mathbf{y}_j)\beta_j\,, \qquad \mathbb{V}\text{ar}\left[u_{h,w}(\mathbf{x}, \cdot)\right] = \sum_{j=1}^Q \beta_j u_h^2(\mathbf{x}, \mathbf{y}_j) - \mathbb{E}\left[u_{h,w}(\mathbf{x}, \cdot)\right]^2.$$

The sequence of sets $\mathcal{I}(w)$, the function $m(i)$ and the family of points to be used at each level characterize the sparse approximation operator $\mathcal{S}_{\mathcal{I}(w)}^m$ introduced in (1.15). The original sparse grid

Approx. space	Collocation: m, g	Galerkin: $\Lambda(w)$
Tensor Product	$m(i) = i$ $g(\mathbf{i}) = \max_n(i_n - 1) \leq w$	$\left\{\mathbf{p} \in \mathbb{N}^N : \ \max_n p_n \leq w\right\}$
Total Degree	$m(i) = i$ $g(\mathbf{i}) = \sum_n(i_n - 1) \leq w$	$\left\{\mathbf{p} \in \mathbb{N}^N : \ \sum_n p_n \leq w\right\}$

Table 1.1: Sparse approximation formulae and corresponding underlying polynomial spaces.

introduced by Smolyak [95] is defined as

$$\mathcal{I}(w) = \left\{\mathbf{i} \in \mathbb{N}^N : \sum_{n=1}^{N}(i_n - 1) \leq w\right\}, \tag{1.18}$$

$$m(i) = \begin{cases} 0 \text{ if } i = 0 \\ 1 \text{ if } i = 1 \\ 2^{i-1} + 1, \text{ if } i > 1, \end{cases} \tag{1.19}$$

and uses nested sequences of points. In Chapter 2 we show the following theorem:

Theorem 1.1. *Given a sequence of polynomial spaces $\mathbb{P}_{\Lambda(w)}(\Gamma)$ such that*

$$\forall \, \mathbf{p} \in \mathbb{P}_{\Lambda(w)}(\Gamma) \quad \mathbf{p} - \mathbf{e}_j \in \mathbb{P}_{\Lambda(w)}(\Gamma) \textit{ for } 1 \leq j \leq N, p_j > 0, \tag{1.20}$$

one can always find a sequence of index sets $\mathcal{I}(w)$ such that the Collocation method delivers approximation in $\mathbb{P}_{\Lambda(w)}(\Gamma)$.

Note that the condition (1.20) on the polynomial space is the counterpart of the admissibility condition (1.16) for the sparse grid. In Chapter 2 we have actually considered sets $\mathcal{I}(w)$ defined as $\mathcal{I}(w) = \left\{\mathbf{i} \in \mathbb{N}^N : g(\mathbf{i}) \leq w\right\}$, with $g : \mathbb{N}^N \to \mathbb{N}$ strictly increasing in each argument, so that (1.20) is automatically satisfied.

Table 1.1 lists the equivalences between the Stochastic Galerkin and Collocation formulation in Tensor Product and Total Degree spaces.

As for quadrature rules, we will mostly use the classical Gaussian rules, see e.g. [85], and the Clenshaw–Curtis rule $y_j^{m(i)} = \cos\left(\frac{(j-1)\pi}{m(i)-1}\right), 1 \leq j \leq m(i)$, see e.g. [102], which results in a nested quadrature rule if used with $m(i)$ as in eq. (1.19).

The rest of this Chapter is devoted to the exposition of the results of the thesis. In particular, we have tried to address the following questions:

1. What is the most effective method in terms of accuracy versus computational cost?
2. What is the best polynomial space $\Lambda(w)$ where the discrete solution $u_{h,w}$ should be shought, again in terms of accuracy versus dimension of the space, and hence computational cost? Or, more generally, can we devise strategies that yield good approximations of the solution with the lowest computational cost possible?

1.2 Results on the regularity of the response surface

A polynomial approximation of the stochastic part of u will be an effective approach if the dependence of u over $\mathbf{y}$ is regular. It is well known (see e.g. [3, 4, 20]) that the solution of (1.7) depends analytically on each parameter $y_n \in \Gamma_n$, under reasonable assumptions on $a(\mathbf{x}, \mathbf{y})$. In particular, denoting $\Gamma_n^* = \prod_{j \neq n} \Gamma_j$ and $\mathbf{y}_n^*$ an arbitrary element of Γ_n^*, there exist regions $\Sigma_n \subset \mathbb{C}$ in the complex plane for $n = 1, \ldots, N$, with $\Sigma_n \supset \Gamma_n$, in which the solution $u(\mathbf{x}, y_n, \mathbf{y}_n^*)$ admits an analytic continuation $u(\mathbf{x}, z, \mathbf{y}_n^*)$, $z \in \Sigma_n$.

However, such results are somehow limited, since the analysis is performed one direction at a time. Instead, we were able to prove the following result, which considers all the y_n at the same time, and also provides a bound on the norms of the derivatives of u, see Chapter 3 for a proof. Note that this result applies to both linear and non-linear expansions like (1.9) and (1.10); a similar result is stated in [21] only for the linear case.

Theorem 1.2. *Let $a(\mathbf{x}, \mathbf{y})$ be a diffusion coefficient for equation (1.5) that satisfies Assumptions 1.1, 1.2 and 1.3. Suppose that $a(\mathbf{x}, \mathbf{y})$ is infinitely many times differentiable with respect to $\mathbf{y}$ and $\exists\, \mathbf{r} \in \mathbb{R}_+^N$ s.t.*

$$\left\| \frac{\partial_{\mathbf{i}} a}{a}(\cdot, \mathbf{y}) \right\|_{L^\infty(D)} \leq \mathbf{r}^{\mathbf{i}} \quad \forall \mathbf{y} \in \Gamma, \tag{1.21}$$

where $\mathbf{i}$ is a multi-index in $\mathbb{N}^N$, $\partial_{\mathbf{i}} a = \dfrac{\partial^{i_1 + \ldots + i_N} a}{\partial y_1^{i_1} \cdots \partial y_N^{i_N}}$, and $\mathbf{r} = (r_1, \ldots, r_N)$ is independent of $\mathbf{y}$, and let $\tilde{\mathbf{r}} = \left(\dfrac{1}{\log 2} \right) \mathbf{r}$. Then

(i) the derivatives of u can be bounded as

$$\|\partial_{\mathbf{i}} u(\mathbf{y})\|_{H_0^1(D)} \leq C_0 |\mathbf{i}|!\, \tilde{\mathbf{r}}^{\mathbf{i}} \quad \forall \mathbf{y} \in \Gamma.$$

with $C_0 = \dfrac{\|f\|_{V'}}{a_{min}}$;

(ii) for every $\mathbf{y}_0 \in \Gamma$ the Taylor series of u converges in the disk

$$\mathcal{D}(\mathbf{y}_0) = \left\{ \mathbf{y} \in \mathbb{R}^N : \tilde{\mathbf{r}} \cdot \mathrm{abs}\,(\mathbf{y} - \mathbf{y}_0) < 1 \right\}.$$

where $\mathrm{abs}\,(\mathbf{v}) = (|v_1|, \ldots, |v_N|)^T$. Therefore $u : \Gamma \to H_0^1(D)$ is analytic and can be extended analytically to the set

$$\Sigma = \left\{ \mathbf{y} \in \mathbb{R}^N : \exists\, \mathbf{y}_0 \in \Gamma \,\text{s.t.}\; \tilde{\mathbf{r}} \cdot \mathrm{abs}\,(\mathbf{y} - \mathbf{y}_0) < 1 \right\}.$$

The same result holds also for u_h, solution of the semidiscrete problem (1.8).

1.3 Results on the Comparison of Stochastic Galerkin and Collocation methods

Quite surprisingly, few works are available in the literature about a full comparison between the Stochastic Galerkin and Collocation methods, see e.g. [30].

As mentioned in Section 1.1.e, in Chapter 2 we show that, given a polynomial space $\mathbb{P}_{\Lambda(w)}(\Gamma)$ such that (1.20) holds, one can always find functions m and g such that the Collocation method delivers approximation in $\mathbb{P}_{\Lambda(w)}(\Gamma)$. This is a very relevant fact to the end of a fair comparison between the two methods, since the two methods obviously should be compared in the same underlying polynomial space.

Once accomplished this, see e.g. Table 1.1, several additional aspects need to be taken into account:

Deegres of freedom: the basis used in the Galerkin method is $\rho(\mathbf{y})d\mathbf{y}$-orthonormal: this implies that the Galerkin method uses the minimum number of degrees of freedom needed to represent the polynomial approximation of $u_{h,w}$ in $V_h(D) \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$, while the number of systems Q to be solved in a Collocation approach can be significantly larger. On the other hand, Collocation deals with uncoupled problems, while the equations in the Galerkin system are coupled.

Parallelization: contrary to the Galerkin method, the Collocation method is trivially parallelizable.

Code reusability the Galerkin method is said to be an "intrusive method", since pre-existing deterministic code can be reused only partially. Collocation can reuse pre-existing code in a "black box" way, although the algorithm may greatly benefit from adjustments to the code aimed to prevent the repetion of identical blocks of operations for each realization.

Implementation issues: in a Galerkin setting a monolitic solver is not feasible and one needs ad-hoc strategies both for the storage and manipulation of the matrix, [82], and for the solution of the linear system, typically variants of preconditioned conjugated gradient, see [83, 32]. Moreover, one may not have analytic formulae for the coefficients of the Galerkin matrix: hence, appropriate (and possibly high dimensional) quadrature rules have to be used to this end, and the consequent approximation error on the entries of the matrix has also to be investigated. On the other hand, assemblying a sparse grid is not an immediate task, but packages are available online and can be coded "once and for all". Among the possible sparse grid codes available we mention the Dakota software developed at SANDIA labs[4] and the Sparse Grid Interpolation Matlab Toolbox developed at Stuttgart University[5]. In this thesis we have developed our own sparse grid Matlab package, that will be made available for download on the internet.

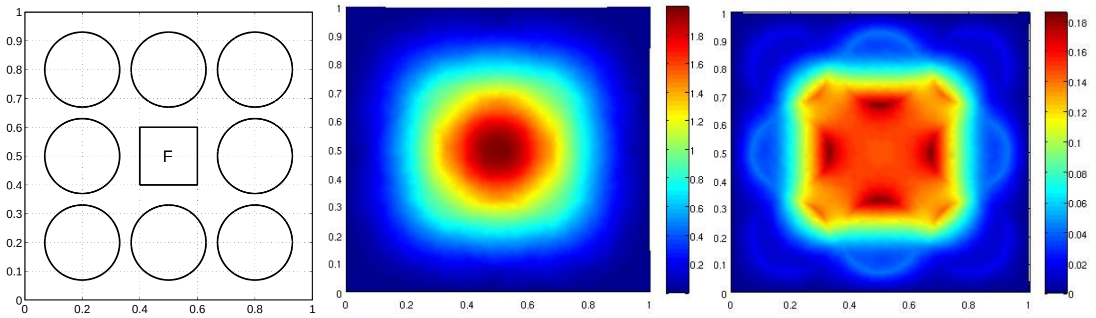
From a more quantitative perspective, in this work we have focused on the comparison of the theoretical computational cost of the Galerkin and Collocation method, and have ingnored instead the real CPU time and setup costs, which greatly depend on the level of optimization of the code used. Such abstract computational cost is defined as the total number of deterministic linear systems solved: this is an obvious definition for the Collocation method, since one has to solve one deterministic problem per collocation point, whereas for the Galerkin method such definition is appropriate as soon as one solves the Galerkin system with the mean-based block-diagonal preconditioned conjugated gradient method proposed in [83]. This choice indeed amounts to the resolution of M deterministic linear systems per PCG iteration, see section 2.3 for details. The costs of the Galerkin and the Collocation solvers will be therefore

$$W_{SG} = \#iter_{PCG} \times M, \qquad W_{SC} = Q. \tag{1.22}$$

respectively. In Chapter 2 we consider the benchmark problem (1.5) with $a(\mathbf{x}, \mathbf{y})$ defined as $a(\mathbf{x}, \mathbf{y}) = 1 + \sum_{i=1}^{8} \chi_i(\mathbf{x})\, y_i$, where $\chi_i(\mathbf{x})$ represent the indicator function of the eight round subdomains depicted in Figure 1.2, and y_i are independent and identically distributed uniform random variables. We refer to section 2.4.a for full discussion on the test case. Here we only show Figure

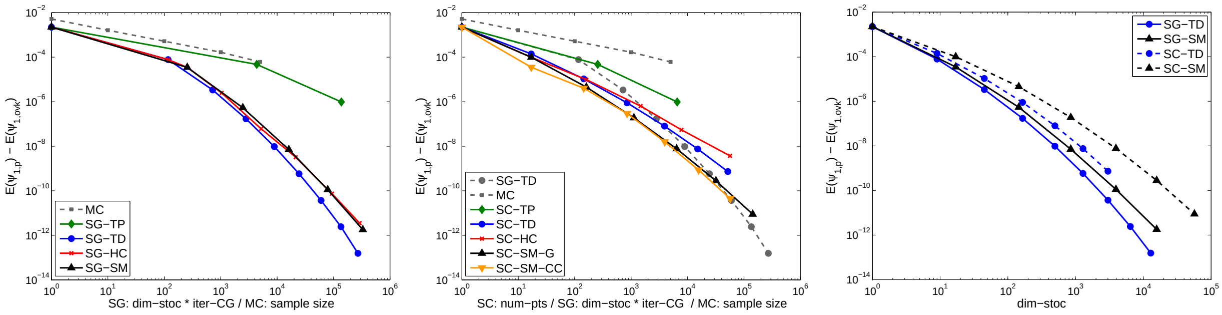
[4] `http://dakota.sandia.gov/index.html`

[5] `http://www.ians.uni-stuttgart.de/spinterp/`



(a) computational domain (b) $\mathbb{E}\left[u(\mathbf{x},\cdot)\right]$ (c) $\sqrt{\mathbb{V}\mathrm{ar}\left[u(\mathbf{x},\cdot)\right]}$

Figure 1.2: results for benchmark case



(a) Galerkin, conv. w.r.t. cost. (b) Collocation, conv. w.r.t. cost. (c) Galerkin and Collocation, conv. w.r.t. dim. of $\mathbb{P}_{\Lambda(w)}(\Gamma)$.

Figure 1.3: Galerkin/Collocation comparison results.

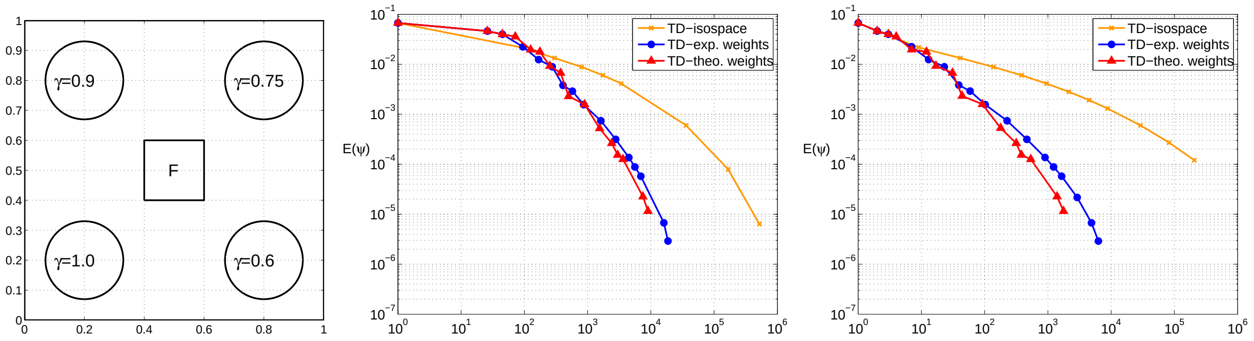
1.3, which is quite representative of the general conclusions of this test. In particular 1.3(a) shows that, as anticipated, in a Galerkin setting the Tensor Product space (cf. Table 1.1) poorly performs due to the excessive costs; all other spaces performe reasonably well and in particular the Total Degree polynomial space is the most effective one. Similarly, Figure 1.3(b) shows that the tensor sparse grid is the worst performing, and the Standard Smolyak Sparse Grid (1.18)-(1.19) is the most convenient one, both using Gauss or Clenshaw–Curtis quadrature points.

Regarding the Galerkin/Collocation comparison, the performances turn out to be really close, with a slight advantage of Collocation for the lower error levels and of Galerkin for the higher ones. Finally Figure 1.3(c) shows that, as expected, the Galerkin method is more effective in terms of error versus dimension of the polynomial space $\mathbb{P}_{\Lambda(w)}(\Gamma)$. Chapter 2 presents several other tests, including one with lognormal random variables, see equation (1.10).

1.4 Optimal Galerkin and Collocation approximations

In the problem considered in the previous Section, all the random variables y_i have the same influence on the solution ("isotropic problem"). This is not the case in a general situation ("anisotropic problem"), so that one should try to adapt the polynomial space $\mathbb{P}_{\Lambda(w)}(\Gamma)$ to the problem at hand, enriching the approximation only with respect to those variables that contribute the most to the variability of the solution.

We can distinguish (at least) two kind of anisotropy strategies: "a-priori" and "a-posteriori". On the one hand, an "a-priori" strategy will try to identify the most suitable polynomial space/sparse



(a) computational domain for anisotropic test

(b) Galerkin, conv. w.r.t. cost

(c) Collocation, conv. w.r.t. cost

Figure 1.4: Convergence results for the anisotropic sets approximation based on theoretical and numerical tuning of the anisotropy weights. See section 2.4.b for more details.

grid through some preliminary analysis, while on the other hand, an "a-posteriori" strategy will adapt the polynomial space/sparse grid as the computation proceeds, typically by some carefully designed "exploration" strategy of the complement of the index sets $\Lambda(w)$ / $\mathcal{I}(w)$.

A number of works are available in literature on "a-posteriori" adaptivity for Sparse Grids (see e.g. [15, 45, 41]), and a few works are available on "a-posteriori" adaptivity for the Galerkin setting (see e.g. [44, 27, 107]). In this work we have instead focused on "a-priori" approaches: we first consider the "anisotropic sets" approach, and then the "optimal sets" approach, which includes the first one as a special case.

1.4.a Anisotropic sets

The first approach is detailed in Chapter 2.4.b, see also [72], where we consider weighted versions of the index sets in Table 1.1, e.g.

$$\Lambda(w) = \left\{ \sum_{n=1}^{N} \alpha_n p_n \leq w \right\}, \qquad \mathcal{I}(w) = \left\{ \sum_{n=1}^{N} \alpha_n (i_n - 1) \leq w \right\} \tag{1.23}$$

and analyze two strategies to tune the anisotropy weights $\boldsymbol{\alpha}$, one theoretical and one experimental, see Chapter 2.4.b for details. We have tested this approach on problem (1.5) where now $a(\mathbf{x}, \mathbf{y}) = 1 + \sum_{i=1}^{4} \gamma_i(\mathbf{x})\, y_i$, where $\gamma_i \in \mathbb{R}$ modify the influence of each y_i on the problem, see Figure 1.4(a). This simple strategy is indeed quite effective, as one can see from the convergence plots in Figure 1.4(b)-1.4(c), even if it completely disregards the interactions among the random variables.

1.4.b Optimal polynomial spaces for Stochastic Galerkin method

To further improve the performances of the anisotropic sets presented in the previous Section, we need to consider a wider class of spaces than (1.23), and look for the most general set $\mathbb{P}_{\Lambda(w)}(\Gamma)$ that maximizes the accuracy given the number of polynomial basis functions M (best M-terms approximation). In other words, our goal is to look for an index set $\mathcal{S} \subset \mathbb{N}^N$ with cardinality M that minimizes the projection error of the gPCE expansion (1.14),

$$\|u - \sum_{\mathbf{p} \in \mathcal{S}} u_{\mathbf{p}} \mathcal{L}_{\mathbf{p}}\|^2_{V \otimes L^2_\rho(\Gamma)} = \sum_{\mathbf{p} \notin \mathcal{S}} \|u_{\mathbf{p}}\|^2_V.$$

The obvious solution to this problem is the set $\mathcal{S}$ that contains the M coefficients $u_{\mathbf{p}}$ with largest norm. This solution of course is not constructive; what we need are sharp estimates of the decay of the coefficients $\|u_{\mathbf{p}}\|_V$, based only on computable quantities, to be used in the approximation of the set $\mathcal{S}$. Seminal works in this direction are [20, 21], where estimates of the decay of the Legendre coefficients are provided.

To fix the ideas, we assume again that a depends on uniform random variables, so that u is expanded in terms of Legendre polynomials. Under the same conditions of Theorem 1.2 it is then possible to prove (see Section 3.3.a) that the following estimate holds for the Legendre coefficients:

$$\|u_{\mathbf{p}}\|_V \leq C_0 e^{-\sum_n g_n p_n} \frac{|\mathbf{p}|!}{\mathbf{p}!}, \qquad g_n = -\log\left(\frac{r_n}{\sqrt{3}\log 2}\right) \tag{1.24}$$

with r_n as in Theorem 1.2. A similar result is given in [21] for the special case $a = a_0 + \sum_{n=1}^N b_n(x) y_n$. We are now in position to define a new sequence $\mathbb{P}_{\Lambda(w)}(\Gamma)$ of polynomial spaces by selecting all multi-indices $\mathbf{p}$ for which the *estimated decay* (1.24) of the corresponding Legendre coefficient is above a fixed threshold ϵ. This in turn corresponds to selecting those indices $\mathbf{p}$ such that

$$\Lambda(w) = \left\{ \mathbf{p} \in \mathbb{N}^N : \sum_{n=1}^N g_n p_n - \log \frac{|\mathbf{p}|!}{\mathbf{p}!} \leq w \right\} \tag{1.25}$$

with $w \in \mathbb{N}^+ = \lceil -\log \epsilon \rceil$. This set closely resembles the anisotropic TD space (1.23), and is therefore called TD-FC set ("TD with factorial correction"). The term $\log \frac{|\mathbf{p}|!}{\mathbf{p}!}$ is indeed a correction that takes into account the intrinsic coupling between the random variables, that is missing in the anisotropic TD sets (1.23) considered in the previous Subsection.

Similarly to the anisotropy weights introduced in the previous Subsection, the quantities $g_n = -\log\left(r_n / \left(\sqrt{3}\log 2\right)\right)$ appearing in (1.25) can be estimated either *a-priori* or numerically, see Section 3.3.a for a full discussion. We remark that estimate (1.24) and set (1.25) are defined for a generic elliptic problem. An soon as one has additional information on the structure of the diffusion coefficient, the optimal set can be further tailored to the problem at hand, see e.g. Sections 3.3.a-3.3.c and Chapter 4.

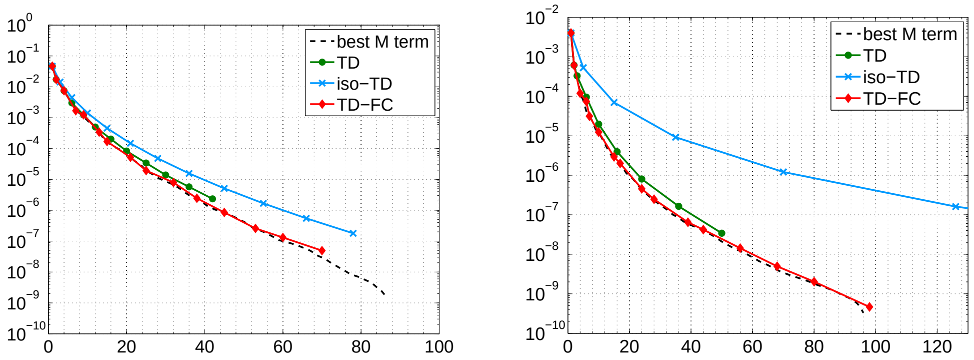
The benchmark test we have considered in this case is the monodimensional version of (1.5),

$$\begin{cases} -(a(x,\mathbf{y})u(x,\mathbf{y})')' = 1 & x \in D = (0,1), \mathbf{y} \in \Gamma \\ u(0,\mathbf{y}) = u(1,\mathbf{y}) = 0, & \mathbf{y} \in \Gamma \end{cases} \tag{1.26}$$

considering a variety of diffusion coefficients $a(x,\mathbf{y})$. Here we report only a few results that confirm the validity of the approach, and refer to Chapter 3 for a wider numerical validation and analysis. Figure 1.5 shows the convergence error for two of the considered diffusion coefficients, for which we can also compute the best M-terms approximation. It is seen that the TD-FC sets show an excellent performance, with a convergence rate essentially equivalent to the best M-terms approximation, confirming that the estimate (1.24) on the decay of the spectral coefficients of u is sharp. The anisotropic TD sets also shows good results, while the remarkably poor performance of the isotropic TD sets confirms the importance of using anisotropic approximations.

1.4.c Optimal Sparse Grids for Stochastic Collocation

We now aim at constructing the optimal sparse grid for Stochastic collocation method, i.e. at choosing the best sequence of index sets $\mathcal{I}(w)$ in equations (1.15)-(1.17). Since the cost of adding a new term to the formula (1.15) heavily depends on the multiindex $\mathbf{i} \in \mathbb{N}^N$ to be added to $\mathcal{I}(w)$,



(a) $a(x, \mathbf{y}) = 1 + 0.1y_1 + 0.5y_2$

(b) $a(x, \mathbf{y}) = 4 + y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$

Figure 1.5: Convergence curves for polynomial approximation with isotropic TD, anisotropic TD and TD-FC polynomial sets. See Section 3.3.b for details on the definition of the error.

we estimate this time the profit $P(\mathbf{i})$ for each index $\mathbf{i} \in \mathbb{N}^N$, and then define the optimal sparse approximation operator $\mathcal{S}^*$ as the one using the set of most profitable indices, i.e.

$$\mathcal{I}^*(\epsilon) = \{\mathbf{i} \in \mathbb{N}_+^N : P(\mathbf{i}) \geq \epsilon\}. \tag{1.27}$$

Following [15, 41] we can define the profit of an index $\mathbf{i}$ as the ratio between its error contribution $\Delta E(\mathbf{i})$ (how much the interpolation error decreases by adding $\mathbf{i}$ to $\mathcal{I}$) and its work contribution $\Delta W(\mathbf{i})$ (how many new grid points appear in the sparse grid):

$$P(\mathbf{i}) = \frac{\Delta E(\mathbf{i})}{\Delta W(\mathbf{i})}, \tag{1.28}$$

and we need to provide computable estimates for these quantities. As for $\Delta W(\mathbf{i})$, it is convenient to use a nested quadrature rule, so that, if $\mathcal{I}$ is admissible, it is possible to compute exactly $\Delta W(\mathbf{i})$ as

$$\Delta W(\mathbf{i}) = \prod_{n=1}^{N} (m(i_n) - m(i_n - 1)). \tag{1.29}$$

As for $\Delta E(\mathbf{i})$, we conjecture that the decay of $\Delta E(\mathbf{i})$ is related to the decay of the gPCE expansion (1.14) of u, through the associated Lebesgue constant

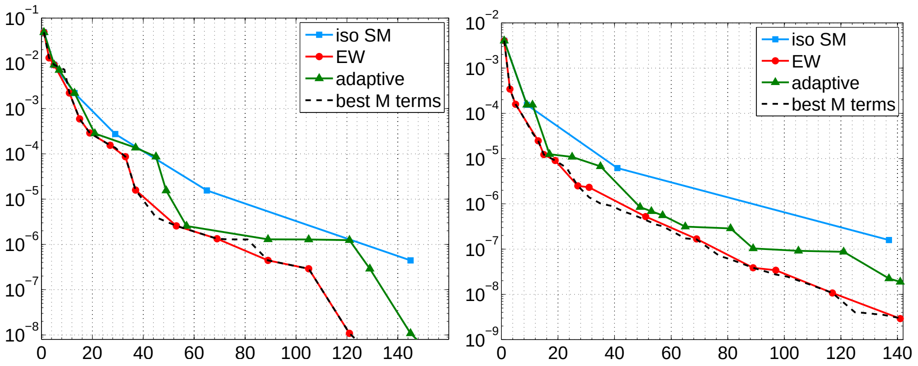
$$\Delta E(\mathbf{i})[u] \lesssim \left\| u_{m(\mathbf{i}-1)} \right\|_V \prod_{n=1}^{N} \mathbb{L}(m(i_n)), \tag{1.30}$$

where $a \lesssim b$ means that there exists a constant c independent of $\mathbf{i}$ such that $a \leq c\, b$. This estimate is reasonable, since it encodes both the information available on the quadrature rules used and on the function u to be interpolated, and indeed turns out to be quite sharp, see Section 3.4.a for details.

We have tested this approach using again problem (1.26), with a depending on N uniform random variables. Using the Clenshaw–Curtis quadrature rule (nested) with $m(i)$ as in eq. (1.19) and Lebesgue constant

$$\mathbb{L}(db(i)) = \frac{2}{\pi}\log(db(i_n) + 1) + 1,$$

and estimate (1.24) for the decay of the gPCE of u depending on uniform random variables, the



(a) $a = 1 + 0.1y_1 + 0.5y_2$

(b) $a(x, \mathbf{y}) = 4 + y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$

Figure 1.6: Performances of the sparse grids based on sets (1.31). Test and error definition are identical to (1.26).

final expression for the optimal sets is

$$\mathcal{I}^*(w) = \left\{ \mathbf{i} \in \mathbb{N}_+^N : \sum_{i=n}^{N} db(i_n - 1) g_n - \log \frac{|db(\mathbf{i} - \mathbf{1})|!}{db(\mathbf{i} - \mathbf{1})!} - \sum_{n=1}^{N} \log \frac{\frac{2}{\pi}\log(db(i_n) + 1) + 1}{db(i_n) - db(i_n - 1)} \leq w \right\}^{ADM} \qquad (1.31)$$

The performance of such grids is shown in Figure 1.6, in which we show the convergence of the standard Smolyak approximation (1.18)-(1.19), the grid computed on the index set containing all multi-indices whose *computed* profit is greater than ϵ ("best M-terms grids") and a grid computed with the version of the "a posteriori" algorithm [41] in the implementation of [56]. More examples can be found in Section 3.4.b.

1.5 An application to groundwater flows

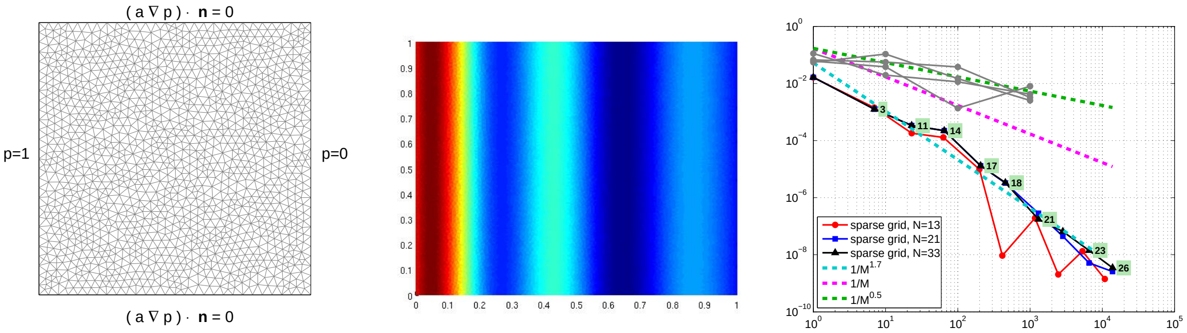
Geophysics is a research area in which Uncertainty Quantification issues naturally arise: whether the focus is on oil reservoir simulation or aquifer management, problems usually deal with such large spatial and temporal time scales that uncertain boundary condition, initial conditions and material properties come often into play.

Within this context, we have focused on the Darcy problem for flows in saturated porous media. Altough specific formulations and discrezations exist for this problem, (see e.g. the mixed-hybrid formulations [13]), we consider here the simple elliptic formulation

$$-\operatorname{div}(a\nabla p) = f \qquad \mathbf{x} \in D = [0,1]^2$$

to be complemented with proper boundary conditions. We focus on a test case where the permeability is random and a pressure gradient induces a flow through two impervious boundaries, see Figure 1.7(a). The random permeability is modeled as a log-normal field, and expanded as in (1.10) with y_i Gaussian random variables. The permeability is assumed to be a stratified material (see Figure 1.7(b)), changing properties along the x_1 axis only, and the logarithm of the permeability is assumed to have a smooth gaussian covariance function, $\exp\left(|x_1 - x_2|^2/L_c^2\right)$, L_c being the correlation length. Such field is then expanded in Fourier series and truncated, thus obtaining

$$\log\left(a(x_1, \mathbf{y}) - \mathbb{E}\left[a(x_1, \mathbf{y})\right]\right) = \sigma\sqrt{c_0}y_0 + \sigma\sum_{k=1}^{K}\sqrt{c_k}\left[\, y_{2k-1}\cos(\omega_k x_1) + y_{2k}\sin(\omega_k x_1)\right], \qquad (1.32)$$



(a) mesh and boundary conditions (b) lognormal permeability realization (c) Monte Carlo and optimal grids convergence

Figure 1.7: Darcy problem with uncertain permeability.

with $y_i \sim \mathcal{N}(0,1)$, $\omega_k = k\pi$ and $c_k \in \mathbb{R}$ (see Section 4.6 for details).

We have considered three different levels of truncation for $\log(a)$ in (1.32): $K = 6, 10, 16$ corresponding to $N = 13, 21, 33$ random variables, so to take into account 99%,99.99% and 100% respectively of the total variability of $\log(a)$. For each truncation we perform a Monte Carlo simulation and compute the sparse grid approximation of the pressure p, specifying the optimal sets (1.27) to the lognormal case, see Section 4.5 for details.

Results are shown in Figure 1.7(c) (see Section 4.5 for details on the computation of the approximation errors). It is seen that all the three sparse grid approximations converge with a rate that is higher than the Monte Carlo one, and moreover seems to be independent of the truncation level. This would mean that the optimal sparse grid construction proposed is quite effective in reducing the deterioration of the performance of the standard sparse grids as the number of random variables increases. Indeed, the selection of the most profitable multiindices manages to "activate " (i.e. to put interpolation points) only in those directions that are most useful in explaining the total variability of the solution, so that the less influent random variables get activated only when the approximation error is sufficiently low. The numbers shown on the convergence plot 1.7(c) indicate the number of random variables activated up to each point. Note also that such procedure allows in principle to work with an infinite number of random variables: the less influent ones will be indeed automatically neglected.

1.6 Proper Generalized Decomposition

The Proper Generalized Decomposition (*PGD* in short) has been introduced to improve the efficiency of the Galerkin method and to overcome some of the issues listed in section 1.3. While a classical Galerkin method fixes "a-priori" the basis for $L^2_\rho(\Gamma)$ as the set of $\rho(\mathbf{y})d\mathbf{y}$-orthonormal polynomials and computes the set of corresponding deterministic modes $\{u_\mathbf{p}\}_{\mathbf{p}\in\Lambda(w)}$, see eq. (1.14), the *PGD* method looks for both the "best" deterministic and the stochastic modes. The solution of (1.11) is written as

$$u^m_{PGD} = \sum_{i=1}^{m} u_i \lambda_i, \qquad u_i \in V_h(D), \quad \lambda_i \in \mathbb{P}_{\Lambda(w)}(\Gamma), \tag{1.33}$$

with u_i, λ_i to be determined, with the goal to approximate the full Galerkin solution (1.14) using $m < M(w)$ modes. To this end, let us consider problem (1.11) in a more compact formulation,

Algorithm 1 Power method

```
$u_{PGD} \leftarrow 0$                                            [element 0 of $\mathcal{V}$]
for $l$ in $1, 2, \dots, m$ do
    Initialize $\lambda$                                          [e.g. at random]
    repeat
        Solve deterministic problem: $u \leftarrow \mathscr{D}(\lambda; u_{PGD})$
        Normalize $u$: $u \leftarrow u/\|u\|_{\mathcal{V}}$
        Solve stochastic problem: $\lambda \leftarrow \mathscr{S}(u; u_{PGD})$
    until $(u, \lambda)$ converged
    $u_{PGD} \leftarrow u_{PGD} + u\lambda$
end for
```

Find $u \in V_h \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$ *such that*

$$A(u,v) = B(v) \quad \forall\, v \in V_h \otimes \mathbb{P}_{\Lambda(w)}(\Gamma), \tag{1.34}$$

with A, B linear with respect to the second argument and first argument, respectively. Suppose now that a *PGD* solution of (1.34) with $m-1$ modes has been computed and that one wants to add a further couple, (u_m, λ_m). This couple is solution of the following Galerkin problem:

Find $(u, \lambda) \in V_h \times \mathbb{P}_{\Lambda(w)}(\Gamma)$ *such that*

$$A\left(u_{PGD}^{m-1} + u\lambda, v\beta\right) = B(v\beta), \qquad \forall (v, \beta) \in V_h \times \mathbb{P}_{\Lambda(w)}(\Gamma)$$

Since u, λ are both unknown, one can then think of computing them with an iterative procedure, solving alternately the two following problems:

Deterministic Problem.

For given $\lambda \in \mathbb{P}_{\Lambda(w)}(\Gamma)$, *find* $u = \mathscr{D}(\lambda; u_{PGD}^{m-1}) \in V_h(D)$ *such that*

$$A\left(u_{PGD}^{m-1} + u\lambda, v\lambda|\pi\right) = B(v\lambda|\pi), \qquad \forall v \in V_h(D). \tag{1.35}$$

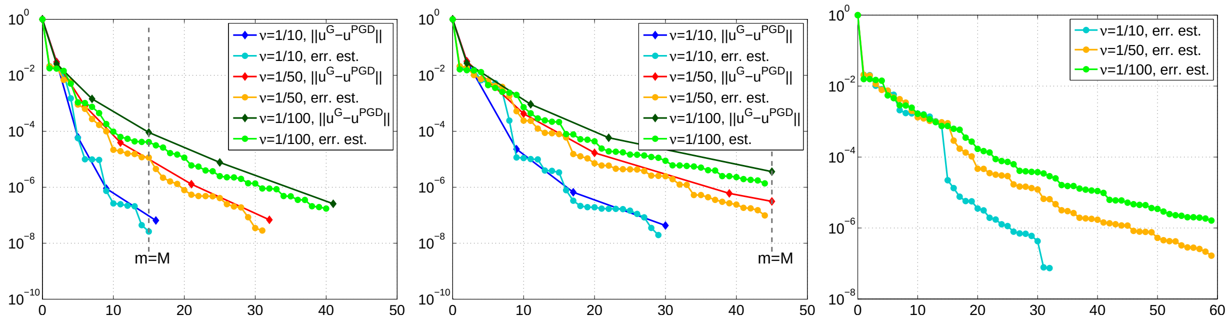
Stochastic Problem.

For given $u \in V_h(D)$, *find* $\lambda = \mathscr{S}(u; u_{PGD}^{m-1}) \in \mathbb{P}_{\Lambda(w)}(\Gamma)$ *such that*

$$A\left(u_{PGD}^{m-1} + u\lambda, u\beta|\pi\right) = B(u\beta|\pi), \qquad \forall \beta \in \mathbb{P}_{\Lambda(w)}(\Gamma). \tag{1.36}$$

Such procedure is called "Power method" (see Algorithm 1), in analogy with the power iteration method for eigenvalues problems. In the case of linear, symmetric, positive definite form A, it was indeed shown in [74, 75] that the sought couples (u, λ) can be interpreted as the solution of a Rayleigh quotient, and that power-iteration like techniques are effective method to compute *PGD* approximations of Galerkin solutions with a reduced computational cost. Their application to scalar non-linear problems has been thoroughly investigated in [78].

We remark that the deterministic problem (1.35) can be indeed solved with minor adaptations to the pre-existing deterministic code, while the stochastic problem (1.36) can be recast as a set of quadratic equations for the coefficients of the orthogonal expansion of $\lambda \in \mathbb{P}_{\Lambda(w)}(\Gamma)$ (see Section 5.3 for details). Moreover, it is important to observe that the total computational cost of the *PGD* method can be remarkably lower than a standard Galerkin solver: computing a Galerkin solution implies the solution of a linear system for $M(w)$ deterministic *coupled* modes, while the



(a) $N = 4$, $M = 15$ (b) $N = 8$, $M = 45$ (c) $N = 15$, $M = 861$

Figure 1.8: error convergence and error estimate with respect to the number of modes m in the PGD solution for $\overline{\nu} = 1/10,\ 1/50,\ 1/100$, for different number of random variables.

PGD solution only requires the resolution of m deterministic *uncoupled* problems, plus the solution of m quadratic equations in $M(w)$ unknowns.

In this Thesis we have further extended the *PGD* procedure, applying it to a non-linear, vector problem, that is the stationary Navier–Stokes equation with uncertain viscosity ν and forcing term (non-linear, vector problem), see Section 5.4 for details. Rather than a simple Power method, we have implemented a *PGD* procedure using the more efficient Arnoldi algorithm, see Section 5.2.

Here we show three test cases, corresponding to problems with $N = 4, 8, 15$ random variables, with viscosity $\nu = 1/10, 1/50, 1/100$. In the first and second case we have computed both the Galerkin solution in the polynomial set $TD(2)$, resulting in $M = 15, 45$ respectively, and its *PGD* approximation, while in the third case we have considered the stochastic polynomial space $TD(3)$ ($M = 816$) and computed only the *PGD* solution. Figure 1.8 shows the convergence of the norm of the approximation error with respect to the number of modes m added in the *PGD* representation, as well as the norm of the stochastic modes λ_i, that can be used as error estimator; for the case $N = 15$, we only show the error estimator. The approximation error correctly decreases as the number of *PGD* modes increases, and in all cases u^m_{PGD} is able to give reasonable approximations of the Galerkin solution with $m \leq M$: this is more and more evident as the number of random variables in the model increases.

Finally, we remark that within this context the pressure reconstruction turns out to be a non-trivial issue: see Section 5.5 for details.

1.7 Global sensitivity analysis for a geocompaction model

Whenever some of the parameters of the problem at hand is uncertain it may be of interest to compute the influence of each of these parameters to the final outcome (i.e. performing a "global sensitiviy analysis"). A possible means of obtaining such information is to compute the so-called Sobol' indices, which derive by a suitable decomposition of the total variance (close to the classical ANOVA sum of squares). In addition, The Sobol' indices can be easily computed from the gPCE expansion of the quantity of interest, which makes them particularly convenient in the framework of the polynomial approximation of PDEs with stochastic coefficients. To introduce the Sobol' indices we first reorder the classical gPCE expansion of a function f depending on N parameters as

$$f(\mathbf{y}) = \sum_{\mathbf{p} \in \mathbb{N}^N} \alpha_{\mathbf{p}} \mathcal{L}_{\mathbf{p}}(\mathbf{y}) = \alpha_{\mathbf{0}} + \sum_{i=1}^{N} \sum_{\mathbf{p} \in \mathcal{P}_i} \alpha_{\mathbf{p}} \mathcal{L}_{\mathbf{p}}(\mathbf{y}) + \sum_{i=1}^{N} \sum_{j=i}^{N} \sum_{\mathbf{p} \in \mathcal{P}_{i,j}} \alpha_{\mathbf{p}} \mathcal{L}_{\mathbf{p}}(\mathbf{y}) + \dots ,$$

where $\mathcal{P}_i$ contains all the multiindices such that only the i-th component is different from 0, $\mathcal{P}_i = \{\mathbf{p} \in \mathbb{N}^N : p_i \neq 0, p_k = 0 \text{ for } k \neq i\}$, and so on. The Sobol' indices can then be computed as:

$$S_{\{i_1,i_2,\dots,i_s\}} = \sum_{\mathbf{p}\in\mathcal{P}_{\{i_1,i_2,\dots,i_s\}}} \frac{\alpha_{\mathbf{p}}^2}{\mathbb{V}\mathrm{ar}\,[f]}, \quad \mathbb{V}\mathrm{ar}\,[f] = \sum_{\mathbf{p}\in\mathbb{N}^N} \alpha_{\mathbf{p}}^2 \ ,$$

and for such indices the following decomposition holds

$$1 = \sum_{i=1}^{N} S_i + \sum_{1\leq i<j\leq N} S_{ij} \ + \ldots + S_{1,2,\dots,N} \,.$$

Hence each Sobol index $S_{\{i_1,i_2,\dots,i_s\}}$ represents the contribution of the corresponding mixed effect to the total variability of f, and it is also easy to compute the total variability due to the i-th random parameter as

$$S_i^T = \sum_{\mathcal{S}_i} S_{\{i_1,i_2,\dots,i_s\}},$$

where the summation is taken over the set $\mathcal{S}_i$ of all index sets $\{\{i_1, i_2, \dots, i_s\}\}$ of any length such that at least one component is i. Thus, a global sensitivity analysis can be easily performed once the gPCE expansion of the quantity of interest is available.

However, if the deterministic solver is very complex (e.g. non linearities, coupled equations, employing iterative solvers ...) it may be not convenient to assemble the Galerkin system to compute the gPCE expansion. To this end, we have developed a novel procedure, that consists in converting a sparse grid approximation of the quantity of interest into a gPCE approximation. This is convenient since the sparse grid approximation is much easier to compute, as it only entails solving a number of independent deterministic problems. The conversion is possible since indeed the sparse grid approximation is a sum of tensor grids, i.e. tensor Lagrangian polynomials, which can be reexpressed as linear combinations of orthogonal polynomials. Note that in general this conversion would require solving as many linear systems as the number of tensor grids which compose the sparse grid. However, if the collocation point chosen are Gaussian quadrature points then no system needs to be solved, since indeed the matrices of the system turn out to be orthogonal.

We have applied this type of analysis to a geocompaction model, that aims at describing the process that transforms sediments into rocks. This process consists of both mechanical stresses (weight of the upper layers, Darcy flows) and chemical reactions, whose action over time reduces the porosity of the sediments, i.e. the empty space among the sediment grains, thus transforming the sediments into rocks.

In our work, we have considered as uncertain the parameters governing the chemical reactions and the mechanical stress. The results of the global sensitivity analysis performed allow to divide the layers in two zones, corresponding to different burial depth: a shallow one in which the compaction process is driven by mechanical actions only, and a deep one in which the compaction process is driven mainly by the chemical reactions. See Chapter 6 for details and results.

Chapter 2

A numerical comparison between Stochastic Spectral Galerkin and Collocation methods

Sections 2.1- 2.4 of this Chapter consist of the paper "J. Bäck, F. Nobile, L. Tamellini, R. Tempone, *Stochastic Spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison*, J.S. Hesthaven and E.M. Ronquist, editors, *Spectral and High Order Methods for Partial Differential Equations*", up to the alignment of the notation and minor improvements in the readibility. In particular, we have adapted the introduction and added some details about the Galerkin matrix (Section 2.3.a), and the numerical setting (Section 2.4).

On the other hand, Section 2.5 is original and contains numerical tests on a problem depending on a set of lognormal random variables.

2.1 Introduction

As anticipated in the introductory part, this chapter deals with the comparison of the Stochastic Galerkin (SG) and Collocation methods (SC) for the computation of statistics of a solution of a PDE with stochastic coefficients. These methods have recently attracted the attention of the Uncertainty Quantification community, since they explore the possible regularity that the solution might have with respect to the input variables, to achieve a convergence rate higher than the one featured by classical sampling methods. The comparison of these two approaches is an open and relevant research topic (see e.g. [30]). This chapter provides, on a couple of numerical examples, a fair comparison between the performances of SG and SC methods *with the same underlying approximation space.*

Traditionally, the SG method approximates the solution in a multivariate polynomial space of given total degree (see e.g. [42, 66, 111] and references therein), or in anisotropic tensor product polynomial spaces [3, 36, 71]. Other global polynomial spaces has been considered recently, see for instance [11, 101], as well as different approximation spaces such as piecewise polynomials [3, 61, 108].

On the other hand the SC method adopted so far for SPDEs follows the classical Smolyak construction, see e.g. [38, 73, 110] and the references therein. It is very relevant to this chapter the fact that the sparse collocation method considered in [110, 73] leads to an approximate solution in a polynomial space, which we call hereafter Smolyak space, that differs from the total degree polynomial space most commonly used in SG approximation.

After having introduced the notation and the setting in Section 2.2, in this chapter we will consider several choices of multivariate polynomial spaces, namely: tensor product (TP), total degree (TD), hyperbolic cross (HC) and Smolyak (SM) spaces. We consider on the one hand, SG approximations in either of these spaces, see Section 2.3.a. On the other hand, we propose a generalization of the classical sparse collocation method that allows us to achieve approximations in these same spaces. By following this path, we are able to compare the two alternative approaches (SG versus SC) given the same underlying multivariate polynomial space, see Section 2.3.b.

Once both SG and SC are posed on the same approximation space the second ingredient in a fair comparison is the computational work associated to each of them for the same level of accuracy, see again Section 2.3. Since SC entails the solution of a number of *uncoupled* deterministic problems, its corresponding computational work is directly proportional to the number of collocation points. On the other hand, SG entails the solution of a large system of *coupled* deterministic problems whose size corresponds to the number of stochastic degrees of freedom (sdof). This can be achieved by an iterative strategy, here chosen to be a Preconditioned Conjugate Gradient solver following [82]. Therefore, a natural approximation of its computational work is given by the product of the number of sdof times the number of iterations performed.

The results of the numerical comparison are shown in Section 2.4. We first present a numerical example having 8 input uniform random variables, in which we compare the performances of the SG and SC methods in terms of accuracy versus (estimated) computational cost. The numerical study shows that the two approaches have comparable performances. Actually, SC seems to be more efficient for errors larger than 10^{-10}, whereas SG is better for smaller errors.

The second numerical example contains 4 input random variables that have largely different influence on the solution. It is thus suited for anisotropic approximations, where higher polynomial degrees are used to discretize the dependence on the random variables that have a greater influence on the solution. We introduce anisotropic versions of both the SG and SC methods and compare their performances for different choices of anisotropy ratios.

Finally, we set up a third test which is identical to the first one but considers lognormal random variables rather than uniform ones, to assess if and to what extent the type of random variables considered affects the performances ot the two methods.

2.2 Problem setting

Let D be a convex bounded polygonal domain in $\mathbb{R}^d$ and $(\Omega, \mathcal{F}, P)$ be a complete probability space. Here Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events and $P : \mathcal{F} \to [0, 1]$ is a probability measure. Consider the stochastic linear elliptic boundary value problem: find a random function, $u : \Omega \times \overline{D} \to \mathbb{R}$, such that P-almost everywhere in Ω, or in other words almost surely (a.s.), the following equation holds:

$$\begin{cases} -\operatorname{div}(a(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)) = f(\mathbf{x}) & \mathbf{x} \in D, \\ u(\mathbf{x},\omega) = 0 & \mathbf{x} \in \partial D. \end{cases} \tag{2.1}$$

where the operators div and ∇ imply differentiation with respect to the physical coordinate only.

The theory presented in this chapter extends straightforwardly to the case of a random forcing term $f = f(\omega, \mathbf{x})$ as well as to a non homogeneous, possibly random, Dirichlet datum on the boundary. For easiness of presentation, we will consider the case where the randomness appears only in the diffusion coefficient, which is, however, the most difficult case, since the solution u depends nonlinearly on it, whereas it depends linearly on the forcing term and boundary data.

For the first part of this chapter, we will specialize Assumptions 1.1, 1.2 and 1.3 as follows:

Assumption 2.1. *$a(\mathbf{x},\omega)$ is strictly positive and bounded with probability 1, i.e. there exist $a_{min} > 0$ and $a_{max} < \infty$ such that*

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

Assumption 2.2. *$a(\mathbf{x},\omega)$ has the form*

$$a(\mathbf{x},\omega) = b_0(\mathbf{x}) + \sum_{n=1}^{N} y_n(\omega) b_n(\mathbf{x}) \tag{2.2}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T : \Omega \to \mathbb{R}^N$, is a vector of independent random variables.

We denote by $\Gamma_n = y_n(\Omega)$ the image set of the random variable y_n, $\Gamma = \Gamma_1 \times \ldots \times \Gamma_N$, and we assume that the random vector $\mathbf{y}$ has a joint probability density function $\rho : \Gamma \to \mathbb{R}_+$ that factorizes as $\rho(\mathbf{y}) = \prod_{n=1}^N \rho_n(y_n)$, $\forall \mathbf{y} \in \Gamma$. Observe that for assumption 2.1 to hold, the image set Γ has to be a bounded set in $\mathbb{R}^N$.

After assumption 2.2, the solution u of (2.1) depends on the single realization $\omega \in \Omega$ only through the value taken by the random vector $\mathbf{y}$. We can therefore replace the probability space $(\Omega, \mathcal{F}, P)$ with $(\Gamma, B(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, where $B(\Gamma)$ denotes the Borel σ-algebra on Γ and $\rho(\mathbf{y})d\mathbf{y}$ is the distribution measure of the vector $\mathbf{y}$.

Finally, we introduce the functional space $H^1(D)$ of square integrable functions in D with square integrable distributional derivatives; its subspace $H^1_0(D)$ of functions with zero trace on the boundary, and the space $L^2_\rho(\Gamma)$ of square integrable functions on Γ with respect to the measure $\rho(\mathbf{y})d\mathbf{y}$. We are now in the position to write a weak formulation of problem (2.1):

Weak Formulation. *Find $u \in H^1_0(D) \otimes L^2_\rho(\Gamma)$ such that $\forall v \in H^1_0(D) \otimes L^2_\rho(\Gamma)$*

$$\int_\Gamma \int_D \left(b_0(\mathbf{x}) + \sum_{n=1}^{N} y_n b_n(\mathbf{x}) \right) \nabla u(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{x},\mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} :$$

$$= \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x},\mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \tag{2.3}$$

Under Assumption 2.1, a straightforward application of the Lax-Milgram lemma allows to prove that there exists a unique solution to problem (2.3) for any $f \in L^2(D)$. Moreover, the following estimate holds:

$$\|\nabla u\|_{L^2(D)\otimes L^2_\rho(\Gamma)} \leq \frac{C_p}{a_{min}} \|f\|_{L^2(D)}$$

where C_p is the Poincaré constant such that $\|u\|_{L^2(D)} \leq C_p \|\nabla u\|_{L^2(D)}$ for any $u \in H^1_0(D)$.

It is well known (see e.g. [4, 71]) that the solution depends analytically on each parameter $y_n \in \Gamma_n$. In particular, denoting $\Gamma^*_n = \prod_{j\neq n} \Gamma_j$ and $\mathbf{y}^*_n$ an arbitrary element of Γ^*_n, there exists a constant M and regions $\Sigma_n \subset \mathbb{C}$ in the complex plane for $n = 1, \ldots, N$, with $\Sigma_n \supset \Gamma_n$, in which the solution $u(\mathbf{x}, y_n, \mathbf{y}^*_n)$ admits an analytic continuation $u(\mathbf{x}, z, \mathbf{y}^*_n)$, $z \in \Sigma_n$. Moreover

$$\max_{z\in\Sigma_n} \max_{\mathbf{y}^*_n \in \Gamma^*_n} \|\nabla u(\cdot, z, \mathbf{y}^*_n)\|_{H^1(D)} \leq M, \qquad \text{for } n = 1, \ldots, N.$$

2.2.a Finite element approximation in the physical space

Let $\mathcal{T}_h$ be a triangulation of the physical domain D and $V_h(D) \subset H_0^1(D)$ a finite element space of piecewise continuous polynomials on $\mathcal{T}_h$, with dimension $N_h = dim(V_h(D))$. We introduce the *semi-discrete* problem:

Weak Formulation. *find* $u_h \in V_h(D) \otimes L^2_\rho(\Gamma)$ *such that* $\forall v_h \in V_h(D)$

$$\int_D \left(b_0(\mathbf{x}) + \sum_{n=1}^{N} y_n b_n(\mathbf{x}) \right) \nabla u_h(\mathbf{x},\mathbf{y}) \cdot \nabla v_h(\mathbf{x})\, d\mathbf{x} = \int_D f(\mathbf{x}) v_h(\mathbf{x})\, d\mathbf{x}, \qquad \rho\text{-}a.e.\ in\ \Gamma. \tag{2.4}$$

Problem (2.4) admits a unique solution for almost every $\mathbf{y} \in \Gamma$. Moreover, u_h satisfies the same analyticity result as the continuous solution u.

Let $\{\phi_i\}_{i=1}^{N_h}$ be a Lagrangian basis of $V_h(D)$ and consider the expansion of the semi-discrete solution as $u_h(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N_h} u_i(\mathbf{y})\phi_i(\mathbf{x})$. Denoting by $\mathbf{U}(\mathbf{y}) = [u_1(\mathbf{y}), \dots, u_{N_h}(\mathbf{y})]^T$ the vector of nodal values as functions of the random variables $\mathbf{y}$, problem (2.4) can be written in algebraic form as

$$\left(K_0 + \sum_{n=1}^{N} y_n K_n \right) \mathbf{U}(\mathbf{y}) = \mathbf{F}, \qquad \rho\text{-a.e. in } \Gamma \tag{2.5}$$

where $(K_n)_{ij} = \int_D b_n(\mathbf{x}) \nabla\phi_j(\mathbf{x}) \cdot \nabla\phi_i(\mathbf{x})$, for $n = 0, \dots, N$, are deterministic stiffness matrices and $\mathbf{F}_i = \int_D f(\mathbf{x})\phi_i(\mathbf{x})$ is a deterministic right hand side.

In writing (2.5) we have heavily exploited the fact that the random diffusion coefficient is an affine function of the random variables y_n. This allows of an efficient evaluation of the stochastic stiffness matrix $A(\mathbf{y}) = K_0 + \sum_{n=1}^{N} y_n K_n$ in any point $\mathbf{y} \in \Gamma$ and greatly simplifies the implementation of the SG method that will be presented in the next section.

2.3 Polynomial approximation in the stochastic dimension

We seek a further approximation of $u_h(\cdot, \mathbf{y})$ with respect to $\mathbf{y}$ by global polynomials, which is sound because of the analyticity of the semi-discrete solution with respect to the input random variables $\mathbf{y}$.

In this chapter we aim at comparing numerically several choices of multivariate polynomials spaces. We remark once more that the choice of the polynomial space is critical when the number of input random variables, N, is large, since the number of stochastic degrees of freedom might grow very fast with N, even exponentially, for instance when isotropic tensor product polynomials are used, cf. (2.6). This effect is known as the *curse of dimensionality.*

Let $w \in \mathbb{N}$ be an integer index denoting the level of approximation and $\mathbf{p} = (p_1, \dots, p_N)$ a multi-index. We introduce a sequence of increasing index sets $\Lambda(w)$ such that $\Lambda(0) = \{(0, \dots, 0)\}$ and $\Lambda(w) \subseteq \Lambda(w+1)$, for $w \geq 0$. Finally, we denote by $\mathbb{P}_{\Lambda(w)}(\Gamma)$ the multivariate polynomial space

$$\mathbb{P}_{\Lambda(w)}(\Gamma) = span \left\{ \prod_{n=1}^{N} y_n^{p_n}, \ \text{ with } \mathbf{p} \in \Lambda(w) \right\}$$

and seek a *fully discrete* approximation $u_{hw} \in V_h(D) \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$. In the following we consider four possible choices of index sets:

Tensor product polynomial space (TP)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \max_{n=1\dots,N} p_n \leq w\} \tag{2.6}$$

Total degree polynomial space (TD)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \sum_{n=1}^{N} p_n \leq w\} \tag{2.7}$$

Hyperbolic cross space (HC)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \prod_{n=1}^{N} (p_n + 1) \leq w + 1\} \tag{2.8}$$

Smolyak polynomial space (SM)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \sum_{n=1}^{n} f_{SM}(p_n) \leq f_{SM}(w)\}, \quad \text{with } f_{SM}(p) = \begin{cases} 0, & p = 0 \\ 1, & p = 1 \\ \lceil \log_2(p) \rceil, & p \geq 2 \end{cases} \tag{2.9}$$

TP and TD spaces are the most common choices. The first suffers greatly from the curse of dimensionality and is impractical for a large dimension N. The second has a reduced curse of dimensionality and has been widely used in SG approximations (see e.g. [42, 66, 77, 97, 111]). HC spaces have been introduced in [2] in the context of approximation of periodic functions by trigonometric polynomials. Recently they have been used to solve elliptic PDEs in high dimension in [93]. Finally, the SM space is an unusual choice in the context of SG approximations. The reason for introducing it will be made clear later, as this space appears naturally when performing interpolation on a sparse grid following the Smolyak construction (see Section 2.3.b). Observe that the Smolyak space is similar to the hyperbolic cross space; indeed, the HC index set can be equivalently written as $\Lambda^{HC}(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \sum_{n=1}^{N} \log_2(p_n+1) \leq \log_2(w+1)\}$. Other polynomial spaces have been introduced e.g. in [101].

It is also useful to introduce *anisotropic* versions of these spaces. Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}_+^N$ be a vector of positive weights, and $\alpha_{min} = \min_n \boldsymbol{\alpha}$. The anisotropic version of the spaces previously defined reads:

Anisotropic tensor product polynomial space (ATP)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \max_{n=1\dots,N} \alpha_n p_n \leq \alpha_{min} w\} \tag{2.10}$$

Anisotropic total degree polynomial space (ATD)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \sum_{n=1}^{N} \alpha_n p_n \leq \alpha_{min} w\} \tag{2.11}$$

Anisotropic hyperbolic cross space (AHC)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \prod_{n=1}^{N} (p_n + 1)^{\frac{\alpha_n}{\alpha_{min}}} \leq w + 1\} \tag{2.12}$$

Anisotropic Smolyak polynomial space (ASM)

$$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \sum_{n=1}^{N} \alpha_n f_{SM}(p_n) \leq \alpha_{min} f_{SM}(w)\} \tag{2.13}$$

In all cases introduced except for the Smolyak space, the maximum polynomial degree used in each direction y_n does not exceed the index w and there is at least one direction (corresponding to the minimum weight α_{min}) for which the monomial y_n^w is in the polynomial space. For the Smolyak space this property holds only if $\log_2(w)$ is integer.

In the next sections we introduce and compare two possible ways of obtaining a *fully-discrete* approximation $u_{h,w} \in V_h(D) \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$, namely Galerkin projection and collocation on a suitable sparse grid.

2.3.a Stochastic Galerkin approximation

The Stochastic Galerkin (SG) - Finite Element approximation consists in restricting the weak formulation (2.3) to the subspace $V_h(D) \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$ and reads:

Weak Formulation. *find* $u_{h,w}^{SG} \in V_h(D) \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$ *such that* $\forall v_{h,w} \in V_h(D) \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$

$$\int_\Gamma \int_D \left(b_0(\mathbf{x}) + \sum_{n=1}^N y_n b_n(\mathbf{x}) \right) \nabla u_{h,w}^{SG}(\mathbf{x},\mathbf{y}) \cdot \nabla v_{h,w}(\mathbf{x},\mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v_{h,w}(\mathbf{x},\mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \quad (2.14)$$

Let $\{L_p\}_{p=0}^\infty$ be the sequence of orthonormal polynomials in Γ_n with respect to the weight ρ_n, i.e. for any $n = 1, \ldots, N$ and $p \geq 0$

$$\int_{\Gamma_n} L_p(t) v(t) \rho_n(t)\, dt = 0 \quad \forall v \in \mathbb{P}_{p-1}(\Gamma_n). \quad (2.15)$$

Given a multi-index $\mathbf{p} = (p_1, \ldots, p_N)$, let $\mathcal{L}_\mathbf{p}(\mathbf{y}) = \prod_{n=1}^N L_{p_n}(y_n)$ be the product of one dimensional orthonormal polynomials. Then a basis for the space $\mathbb{P}_{\Lambda(w)}(\Gamma)$ is given by $\{\mathcal{L}_\mathbf{p}\}_{\mathbf{p}\in\Lambda(w)}$ and the SG solution can be expanded as

$$u_{h,w}^{SG}(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{p}\in\Lambda(w)} u_\mathbf{p}(\mathbf{x}) \mathcal{L}_\mathbf{p}(\mathbf{y}) = \sum_{\mathbf{p}\in\Lambda(w)} \sum_{i=1}^{N_h} u_{\mathbf{p},i} \phi_i(\mathbf{x}) \mathcal{L}_\mathbf{p}(\mathbf{y}). \quad (2.16)$$

Given this expansion and exploiting the orthonormality of the basis $\{\mathcal{L}_\mathbf{p}(\mathbf{y})\}_{\mathbf{p}\in\Lambda(w)}$, one can easily compute mean and variance of $u_{h,w}^{SG}$ as $\mathbb{E}\left[u_{h,w}^{SG}\right](\mathbf{x}) = u_\mathbf{0}(\mathbf{x})$ and $\mathbb{V}\mathrm{ar}\left[u_{h,w}^{SG}\right](\mathbf{x}) = \sum_{\mathbf{p}\in\Lambda(w)} u_\mathbf{p}^2(\mathbf{x}) - \mathbb{E}\left[u_{h,w}^{SG}\right]^2(\mathbf{x})$.

Let $\mathbf{U_p} = [u_{\mathbf{p},1}, \ldots, u_{\mathbf{p},N_h}]^T$ be the vector of nodal values of the finite element solution corresponding to the $\mathbf{p}$ multi-index. Then inserting expression (2.16) into (2.14) and recalling the definition of the deterministic stiffness matrices K_n, we obtain the *system of* $M = dim(\mathbb{P}_{\Lambda(w)}(\Gamma))$ *coupled finite element problems*

$$K_0 \mathbf{U_p} + \sum_{n=1}^N \sum_{\mathbf{q}\in\Lambda(w)} G_{\mathbf{p},\mathbf{q}}^n K_n \mathbf{U_q} = \mathbf{F}\delta_{\mathbf{0p}}, \qquad \forall \mathbf{p} \in \Lambda(w). \quad (2.17)$$

where

$$G_{\mathbf{p},\mathbf{q}}^n = \int_\Gamma y_n \mathcal{L}_\mathbf{p}(\mathbf{y}) \mathcal{L}_\mathbf{q}(\mathbf{y}) \rho(\mathbf{y})\, d\mathbf{y} \quad (2.18)$$

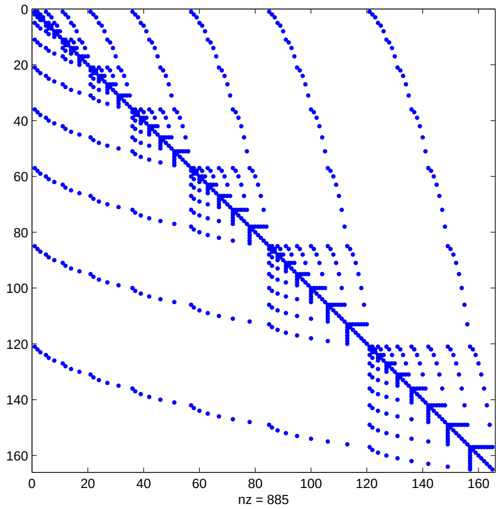


Figure 2.1: sparsity plot for the SG matrix. Here we consider $\Lambda(w)$ =TD(3).

and δ_{ij} is the usual Kroneker symbol. $G^n_{\mathbf{p},\mathbf{q}}$ can be explicitly calculated via the well-known three-terms relation for orthogonal polynomials, see e.g. [39, 85].

The resulting matrix of the algebraic system (2.17) is highly sparse, see Figure 2.1, symmetric and positive definite. For its solution we consider a Preconditioned Conjugate Gradient (PCG) method with block diagonal preconditioner

$$P_{\mathbf{q},\mathbf{q}} = K_0 + \sum_{n=1}^{N} G^n_{\mathbf{q},\mathbf{q}} K^n \tag{2.19}$$

as suggested in [82]. It follows easily from Assumption 2.1 that the condition number of the preconditioned matrix is independent of the discretization parameters both in the physical and stochastic spaces, and therefore the preconditioner is optimal. See [32, 83] for a detailed analysis of the condition number of the SG matrix.

Each PCG iteration implies the solution of M deterministic problems with matrix $P_{\mathbf{q},\mathbf{q}}$. If the finite element discretization is relatively coarse and the dimension of the probability space is moderate, a Cholesky factorization of all matrices $P_{\mathbf{q},\mathbf{q}}$ could be computed once and for all. In general, this strategy could lead to excessive memory requirements and an iterative method should be preferred. Observe that in certain cases (e.g. for uniform random variables) all blocks are equal and this reduces considerably the computational burden.

Let us now denote by W_{FE} the cost for solving one deterministic problem and by N_{iter} the number of PCG iterations. In this chapter we focus on the computational cost for solving the linear system (2.17) and neglect the time for assembling the full stochastic matrix, which highly depends on how much the computer code has been optimized. Therefore, we can estimate the total cost W_{SGFE} for SG - finite element as

$$W_{SGFE} \approx M \times W_{FE} \times N_{iter}. \tag{2.20}$$

This estimate will be used to compare the SG method with the SC method in the numerical tests presented in Section 2.4.

2.3.b Stochastic collocation approximation on sparse grids

The Stochastic Collocation (SC) - Finite Element method consists in collocating the semi-discrete problem (2.4) in a set of points $\{\mathbf{y}_j \in \Gamma, \ \ j = 1, \dots, Q\}$, i.e. computing the solutions $u_h(\cdot, \mathbf{y}_j)$ and building a global polynomial approximation $u^{SC}_{h,w}$ (not necessarily interpolatory) upon those evaluations: $u^{SC}_{h,w}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{Q} u_h(\mathbf{x}, \mathbf{y}_j)\mathscr{P}_j(\mathbf{y})$ for suitable multivariate polynomials $\{\mathscr{P}_j\}_{j=1}^{Q}$.

We consider here a generalization of the classical Smolyak construction (see e.g. [95, 7]) to build a multivariate polynomial approximation on a sparse grid. For each direction y_n we introduce a sequence of one dimensional polynomial interpolant operators of increasing order: $\mathcal{U}_n^{m(i)} : C^0(\Gamma_n) \to \mathbb{P}_{m(i)-1}(\Gamma_n)$. Here $i \geq 1$ denotes the level of approximation and $m(i)$ the number of collocation points used to build the interpolation at level i, with the requirement that $m(1) = 1$ and $m(i) < m(i+1)$ for $i \geq 1$. In addition, let $m(0) = 0$ and $\mathcal{U}_n^0 = 0$. In this chapter the collocation points $\{y^{(i)}_{n,j}, \ j = 1, \dots, m(i)\}$ for the one dimensional interpolation formula $\mathcal{U}_n^{m(i)}$ will be taken as the Gauss points with respect to the weight ρ_n, that is the zeros of the orthogonal polynomial $L_{m(i)}$ defined in (2.15). To simplify the presentation of the sparse grid approximation (2.21), we now introduce the difference operators

$$\Delta_n^{m(i)} = \mathcal{U}_n^{m(i)} - \mathcal{U}_n^{m(i-1)}.$$

Given an integer $w \geq 0$ and a multi-index $\mathbf{i} = (i_1, \dots, i_N) \in \mathbb{N}^N_+$, $\mathbf{i} \geq \mathbf{1}$, we introduce a function $g : \mathbb{N}^N_+ \to \mathbb{N}$ strictly increasing in each argument and define a sparse grid approximation of u_h as

$$u^{SC}_{h,w} = \mathcal{S}^m_{\mathcal{I}(w)}[u_h] = \sum_{\mathcal{I}(w)} \bigotimes_{n=1}^{N} \Delta_n^{m(i_n)}(u_h), \qquad \mathcal{I}(w) = \{\mathbf{i} \in \mathbb{N}^N_+ : g(\mathbf{i}) \leq w\}. \tag{2.21}$$

The previous formula implies evaluation of the function u_h in a finite set of points $\mathcal{H}^m_{\mathcal{I}(w)} \subset \Gamma$ (*sparse grid*). From the construction (2.21) one can easily build the corresponding quadrature formula, and evaluate e.g.

$$\mathbb{E}\left[u^{SC}_{h,w}\right](\mathbf{x}) = \sum_{j=1}^{Q} \beta_j u_h(\mathbf{x}, \mathbf{y}_j), \qquad \mathbb{V}\text{ar}\left[u^{SC}_{h,w}\right] = \sum_{j=1}^{Q} \beta_j u_h^2(\mathbf{x}, \mathbf{y}_j) - \mathbb{E}\left[u^{SC}_{h,w}\right]^2(\mathbf{x}).$$

To fully characterize the sparse approximation operator $\mathcal{S}^m_{\mathcal{I}(\omega)}$ one has to provide the two strictly increasing functions $m : \mathbb{N}_+ \to \mathbb{N}_+$ and $g : \mathbb{N}^N_+ \to \mathbb{N}$. The first defines the relation between the level i and the number of points $m(i)$ in the corresponding one dimensional polynomial interpolation formula $\mathcal{U}^{m(i)}$, while the second characterizes the set of multi-indices used to construct the sparse approximation. Since m is not surjective in $\mathbb{N}^+$ (unless it is affine) we introduce a *left inverse* $m^{-1}(k) = \min\{i \in \mathbb{N}_+ : \ m(i) \geq k\}$. Observe that with this choice m^{-1} is a (non-strictly) increasing function satisfying $m^{-1}(m(i)) = i$, and $m(m^{-1}(k)) \geq k$.

Let $\mathbf{m}(\mathbf{i}) = (m(i_1), \dots, m(i_N))$ and consider the polynomial order set

$$\Lambda^{m,g}(w) = \{\mathbf{p} \in \mathbb{N}^N, \ \ g(\mathbf{m}^{-1}(\mathbf{p} + \mathbf{1})) \leq w\}.$$

The following result characterizes the polynomial space underlying the sparse approximation $\mathcal{S}^m_{\mathcal{I}(w)}[u_h]$:

Proposition 2.1.

a) For any $f \in C^0(\Gamma)$, we have $\mathcal{S}^m_{\mathcal{I}(w)}[f] \in \mathbb{P}_{\Lambda^{m,g}(w)}$.

b) Moreover, $\mathcal{S}^m_{\mathcal{I}(w)}[v] = v, \;\; \forall v \in \mathbb{P}_{\Lambda^{m,g}(w)}$*.*

Proof. Let us denote by $\mathbb{P}_{\mathbf{m(i)-1}}$ the tensor product polynomial space

$$\mathbb{P}_{\mathbf{m(i)-1}} = span\left\{ \prod_{n=1}^{N} y_n^{p_n}, \;\; p_n \leq m(i_n) - 1 \right\}.$$

Clearly we have that $\bigotimes_{n=1}^{N} \Delta_n^{m(i_n)}(f) \in \mathbb{P}_{\mathbf{m(i)-1}}(\Gamma)$ and

$$\begin{aligned}
\mathcal{S}^m_{\mathcal{I}(w)}[f] \in span&\Big\{ \bigcup_{\mathbf{i}\in\mathbb{N}^N_+ :\, g(\mathbf{i})\leq w} \mathbb{P}_{\mathbf{m(i)-1}}(\Gamma) \Big\} \\
&\equiv span\Big\{ \bigcup_{\mathbf{i}\in\mathbb{N}^N_+ :\, g(\mathbf{i})\leq w} span\{\prod_{n=1}^{N} y_n^{p_n}, \;\; \mathbf{p} \leq \mathbf{m(i)-1}\}\Big\} \\
&\equiv span\Big\{ \bigcup_{\mathbf{i}\in\mathbb{N}^N_+ :\, g(\mathbf{i})\leq w} span\{\prod_{n=1}^{N} y_n^{p_n}, \;\; \mathbf{m}^{-1}(\mathbf{p+1}) \leq \mathbf{i}\}\Big\} \\
&\equiv span\{\prod_{n=1}^{N} y_n^{p_n}, \;\; g(\mathbf{m}^{-1}(\mathbf{p+1})) \leq w\} =: \mathbb{P}_{\Lambda^{m,g}(w)}(\Gamma).
\end{aligned}$$

This proves a). Due to linearity in (2.21), to prove point b) we only need to show that the approximation formula $\mathcal{S}^m_{\mathcal{I}(w)}$ is exact for all monomials $\prod_{n=1}^{N} y_n^{p_n}$ with $\mathbf{p} \in \Lambda^{m,g}(w)$. We have

$$\begin{aligned}
\mathcal{S}^m_{\mathcal{I}(w)}\left[\prod_{n=1}^{N} y_n^{p_n}\right] &= \sum_{\mathcal{I}(w)} \bigotimes_{n=1}^{N} \Delta_n^{m(i_n)} \mathbf{y^p} \\
&= \sum_{\mathcal{I}(w)} \prod_{n=1}^{N} \left((\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)}) y_n^{p_n} \right).
\end{aligned}$$

Observe that $\mathcal{U}^{m(i_n)} y_n^{p_n}$ will be an exact interpolation whenever $m(i_n) \geq p_n + 1$ and therefore the term $\prod_{n=1}^{N} (\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)}) y_n^{p_n}$ will vanish if any of the $m(i_n - 1) \geq p_n + 1$ or equivalently if there exists at least one n such that $i_n \geq m^{-1}(p_n + 1) + 1$. Let $\bar{i}_n = m^{-1}(p_n + 1)$ for $n = 1, \ldots, N$. The multi-index $\bar{\mathbf{i}} = (\bar{i}_1, \ldots, \bar{i}_N)$ satisfies the constraint $g(\bar{\mathbf{i}}) \leq p$.

Then, the previous formula reduces to

$$\begin{aligned}
\mathcal{S}^m_{\mathcal{I}(w)}\left[\prod_{n=1}^{N} y_n^{p_n}\right] &= \sum_{\mathbf{i}\leq\bar{\mathbf{i}}} \prod_{n=1}^{N} \left((\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)}) y_n^{p_n} \right) \\
&= \prod_{n=1}^{N} \sum_{i_n=0}^{\bar{i}_n} \left((\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)}) y_n^{p_n} \right) = \prod_{n=1}^{N} \mathcal{U}^{m(\bar{i}_n)} y_n^{p_n}.
\end{aligned}$$

The final result follows from the fact that $m(\bar{i}_n) = m(m^{-1}(p_n + 1)) \geq p_n + 1$ and therefore the interpolant $\mathcal{U}^{m(\bar{i}_n)}$ is exact for $y_n^{p_n}$.

□

Remark 2.1. *Observe that in the previous Lemma we have never used the Assumption that the one dimensional interpolants are based on Gauss points. Hence, the previous result still holds for interpolants based on arbitrary (distinct) knots and for an arbitrary strictly increasing function* $m(i)$*.*

We recall that the most typical choice of m and g is given by (see [7, 95])

$$m(i) = \begin{cases} 1, & \text{for } i = 1 \\ 2^{i-1}+1, & \text{for } i > 1 \end{cases} \text{ and } g(\mathbf{i}) = \sum_{n=1}^{N} (i_n - 1).$$

This choice of m, combined with the choice of Clenshaw-Curtis interpolation points (extrema of Chebyshev polynomials) leads to nested sequences of one dimensional interpolation formulas and a reduced sparse grid. In the same vein, it is possible to show that the underlying polynomial space associated to the operator $S^m_{\mathcal{I}(w)}$ is the Smolyak space $\mathbb{P}_{\Lambda(w)}$ defined in (2.9).

On the other hand, if we choose $m(i) = i$, it is easy to find functions g for the construction of sparse collocation approximations in the polynomial spaces introduced in Section 2.3, namely tensor product (2.6), total degree (2.7) and hyperbolic cross (2.8) spaces. Table 2.1 summarizes several available. It is also straightforward to build the corresponding anisotropic sparse approximation formulas.

Approx. space	Collocation: m, g	Galerkin: $\Lambda(w)$
Tensor Product	$m(i) = i$ $g(\mathbf{i}) = \max_n (i_n - 1) \leq w$	$\{\mathbf{p} \in \mathbb{N}^N : \ \max_n p_n \leq w\}$
Total Degree	$m(i) = i$ $g(\mathbf{i}) = \sum_n (i_n - 1) \leq w$	$\{\mathbf{p} \in \mathbb{N}^N : \ \sum_n p_n \leq w\}$
Hyperbolic Cross	$m(i) = i$ $g(\mathbf{i}) = \prod_n (i_n) \leq w + 1$	$\{\mathbf{p} \in \mathbb{N}^N : \ \prod_n (p_n + 1) \leq w + 1\}$
Smolyak	$m(i) = \begin{cases} 2^{i-1}+1, \ i > 1 \\ 1, \ i = 1 \end{cases}$ $g(\mathbf{i}) = \sum_n (i_n - 1) \leq w$	$\{\mathbf{p} \in \mathbb{N}^N : \ \sum_n f_{SM}(p_n) \leq f_{SM}(w)\}$

Table 2.1: Sparse approximation formulas and corresponding underlying polynomial space

Let now $\mathcal{H}^m_{\mathcal{I}(w)}$ be the sparse grid associated to the formula $\mathcal{S}^m_{\mathcal{I}(w)}$ and Q the number of distinct collocation points in $\mathcal{H}^m_{\mathcal{I}(w)}$. To form the sparse collocation solution $u_{h,w}$ we only have to solve Q *independent* deterministic problems. Observe, however, that in general the number of points Q is much larger than the dimension M of the corresponding polynomial space $\mathbb{P}_{\Lambda^{m,g}(w)}$. The computational cost of the SC - Finite Element method can therefore be estimated as

$$W_{SCFE} \approx Q \times W_{FE}, \tag{2.22}$$

to be compared with the cost of the SG - Finite Element method in the same polynomial space, given by (2.20).

2.4 Numerical results

2.4.a Test case 1: isotropic problem

In this first test case we consider a thermal diffusion problem in the form of (2.1) defined in the unit square $[0, 1]^2$, with homogeneous Dirichlet boundary conditions and stochastic conductivity coefficient that depends on a finite, small, number of random variables. The coefficient is chosen in

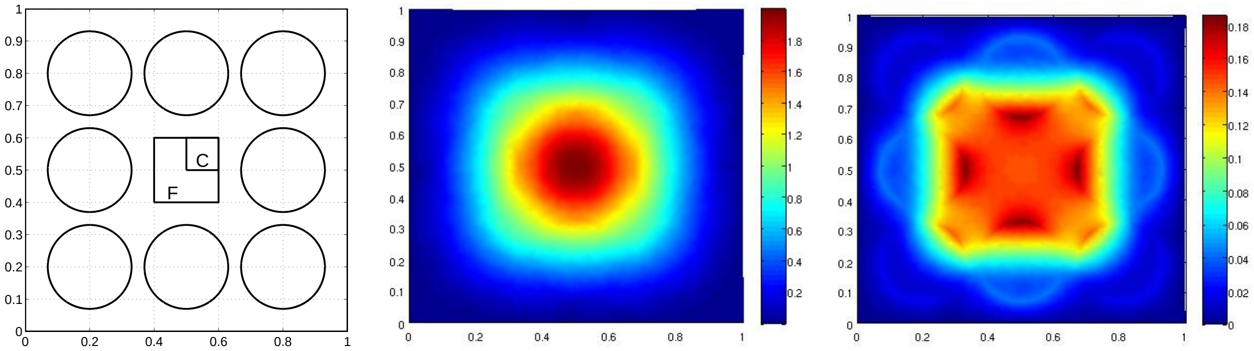


Figure 2.2: Left: geometry for test case 1. Middle: expected value of the solution. Right: standard deviation of the solution.

such a way that each random input has more or less the same influence on the solution (isotropic problem).

Fig. 2.2-left shows the geometry of the test case. The forcing term is deterministic, $f(\mathbf{x}) = 100\chi_F(\mathbf{x})$, where $\chi_F(\mathbf{x})$ is the indicator function of F, a square subdomain with side length equal to 0.2, centered in the domain. The material features 8 circular inclusions with radius $r = 0.13$ and symmetrically distributed with respect to the center of the square, each with a uniformly distributed random conductivity. Let $\chi_n(\mathbf{x}), n = 1, .., 8$ be the indicator function for each circle. The expression of the stochastic conductivity coefficient is then in the form of (2.2), with $b_n(\mathbf{x}) = \chi_n(\mathbf{x})$:

$$a(\mathbf{x},\omega) = b_0(\mathbf{x}) + \sum_{n=1}^{8} \chi_n(\mathbf{x}) y_n(\omega), \quad \text{with } b_0 = 1 \text{ and } y_n(\omega) \sim \mathcal{U}(y_{min}, y_{max}), \tag{2.23}$$

with $y_{min} = -0.99$, $y_{max} = -0.2$. As a consequence, the basis functions $\mathcal{L}_{\mathbf{p}}(\mathbf{y})$ for SG methods will be Legendre polynomials orthonormal with respect to the uniform probability measure in $[y_{min}, y_{max}]$, and the collocation points for SC will be the corresponding Gauss points. Using the orthogonality property and the three-terms relation for the Legendre polynomials, it is easy to see from (2.18) that for an extra-diagonal block in (2.17) to be nonzero the following condition must hold

$$\exists\, \overline{n} : \quad |p_{\overline{n}} - q_{\overline{n}}| = 1, \;\; p_j = q_j \;\; \forall j \neq \overline{n}, \tag{2.24}$$

is which case only the $\overline{n}$-th random variable will contribute, i.e. only $G^{\overline{n}}_{\mathbf{p},\mathbf{q}}$ will be nonzero. On the main diagonal instead all the random variables will contribute because of the rescaling of the polynomials in $[y_{min}, y_{max}]$, and such contribution will be indentical for all $\mathbf{q} \in \Lambda(w)$,

$$G^{n}_{\mathbf{q},\mathbf{q}} = \frac{y_{min} + y_{max}}{2} \qquad \forall n = 1, \ldots, 8, \quad \forall \mathbf{q} \in \Lambda(w). \tag{2.25}$$

Therefore, the preconditioner (2.19) will also be the same for all $\mathbf{q} \in \Lambda(w)$,

$$P_{\mathbf{q},\mathbf{q}} = K_0 + \frac{y_{min} + y_{max}}{2} \sum_{n=1}^{N} K_n \, . \tag{2.26}$$

We will compare the accuracy of the Stochastic Galerkin (SG) and Stochastic Collocation (SC) methods by looking at statistical indicators of two quantities of interest:

- $\psi_1(u) = \int_F u(\mathbf{x}) d\mathbf{x}$;

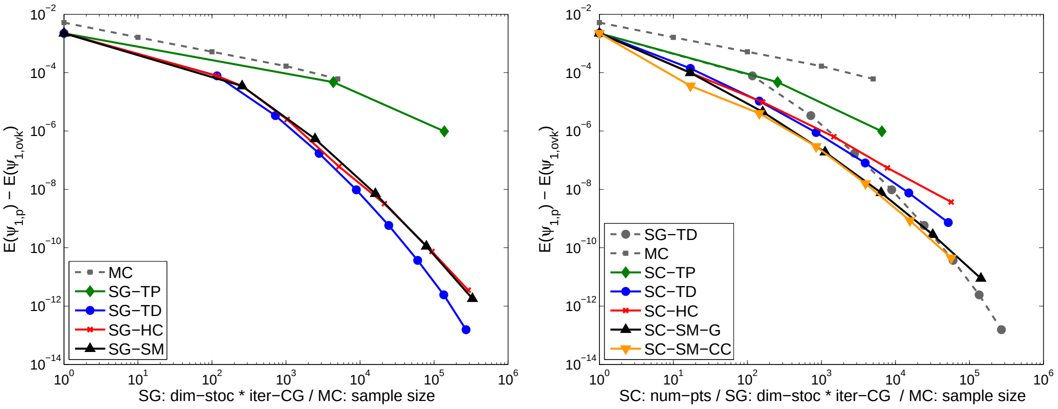


Figure 2.3: Error $\varepsilon_{\text{mean}}[\psi_1]$ versus estimated computational cost. Left: comparison between SG methods and Monte Carlo. Right: comparison between SC methods and SG-TD.

- $\psi_2(u) = \int_C \partial_x u(\mathbf{x}) d\mathbf{x}$.

The quantity $\psi_2(u)$ is defined only on C, the upper right part of F, since by symmetry its expected value on F is 0 whatever (isotropic) Galerkin or Collocation approximation is considered.

Let u_p be an approximate solution (computed either with SG or SC) and u_{ex} the exact solution. For both quantities ψ_1 and ψ_2 we will check the convergence of the following errors:

- error in the mean: $\varepsilon_{\text{mean}}[\psi_j] = |\mathbb{E}[\psi_j(u_p)] - \mathbb{E}[\psi_j(u_{ex})]|$;
- error in the variance: $\varepsilon_{\text{var}}[\psi_j] = |\mathbb{V}\text{ar}[\psi_j(u_p)] - \mathbb{V}\text{ar}[\psi_j(u_{ex})]|$;
- error in L^2 norm: $\varepsilon_{\text{norm}}[\psi_j] = \sqrt{\mathbb{E}[(\psi_i(u_p) - \psi_i(u_{ex}))^2]}$.

Since we do not know the exact solution for this problem, we will check the convergence of the statistical indicators with respect to an overkill solution, which we consider close enough to the exact one. To this end we take the solution computed with SG-TD at level 9, which has approximately 24000 stochastic degrees of freedom (sdof). The L^2 error will be calculated via a Monte Carlo approximation, i.e.

$$\varepsilon_{\text{norm}}[\psi_j] \simeq \frac{1}{Q_{MC}} \left(\sum_{l=1}^{Q_{MC}} [\psi_j(u_p(\mathbf{y}_l)) - \psi_j(u_{ex}(\mathbf{y}_l))]^2 \right)^{1/2},$$

where $\mathbf{y}_l$, $l = 1, .., Q_{MC}$, are randomly chosen points in Γ. To this end we have used $Q_{MC} = 1000$ points.

We remark that here and in the following test all the computations are performed on the same physical mesh, which is supposed to be refined enough to solve adequately the elliptic problem for every value $\mathbf{y}$ of the random variables. Moreover notice that, as stated in section 2.1, the FEM solution and the exact solution have the same regularity with respect to the stochastic variables. Therefore we expect the convergence in the stochastic dimension not to be affected by space discretization.

We have compared the performances of the SG and Collocation methods with the four choices of polynomial spaces presented in Table 2.1. In our convergence plots we have also added the performance of the classical Monte Carlo method.

Fig. 2.3 shows the error $\varepsilon_{\text{mean}}[\psi_1]$ versus the estimated computational cost (normalized to the cost W_{FE} of a deterministic solve) given by formula (2.20) for SG methods and (2.22) for SC

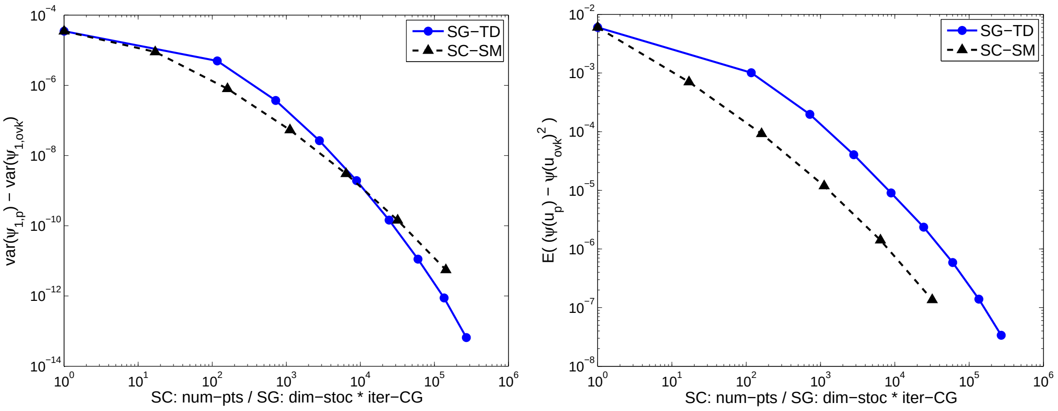


Figure 2.4: Convergence curves for $\varepsilon_{\text{var}}\left[\psi_1\right]$ (left) and $\varepsilon_{\text{norm}}\left[\psi_1\right]$ (right) with respect to the computational cost. Comparison between SG-TD and SC-SM methods.

methods. For the Monte Carlo method the cost is simply $M \times W_{FE}$, where M is the number of samples used. The Monte Carlo has been repeated 20 times and only the average error over the 20 repetitions is shown.

As one can see, Monte Carlo has the worst performance, followed by tensor product polynomial spaces both in the SG and SC version, as expected. All other choices lead to similar, however much more accurate, results, with TD being the best space for Galerkin method and SM the best for Collocation.

We notice that different choices of collocation points for SC-SM (Gauss versus Clenshaw Curtis) lead to similar results (see Fig. 2.3-right). Therefore from now on we will only use SC-SM with Gauss points.

From Fig. 2.3-right we conclude that the SC method is the best method with respect to the computational cost, at least for “practical” tolerances, while, for very small tolerances ($\leq 10^{-10}$), SG is a better choice. The same happens also for the other error indicators $\varepsilon_{\text{var}}\left[\psi_1\right]$ and $\varepsilon_{\text{norm}}\left[\psi_1\right]$, (see Fig. 2.4), as well as for the quantity ψ_2 (see Fig. 2.5).

We should point out that the plots may not represent a completely fair comparison. Actually, the solution of the global linear system for SG method is performed through preconditioned conjugate gradient iterations, with a fixed tolerance ($\epsilon = 10^{-12}$); this clearly over-resolves the system when the error in the stochastic dimension is much larger than ϵ. The performance of SG may be therefore improved by tuning the tolerance of the PCG method to an *a posteriori* estimation of the stochastic error. However, we have observed that running the same SG simulations with tolerance $\epsilon = 10^{-8}$ changes only slightly the results, so we can say that the choice of the tolerance for the PCG method is not deeply affecting our performance/cost analysis.

It is also instructive to look at the convergence plots of the error versus the dimension of the stochastic space (Fig. 2.6). As expected from L^2 optimality, for a given polynomial space the Galerkin solution is more accurate than the collocation solution. We remind once more, however, that the computational cost in the two cases is quite different and the convergence plots in Fig. 2.3 give a more complete picture of the performances of the two methods.

2.4.b Test case 2: anisotropic problem

In this test we consider an anisotropic problem in which different random variables contribute differently to the total variability of the solution, in order to study the advantages of the anisotropic version of the SC and SG methods. We take the geometry and problem definition similar to test case

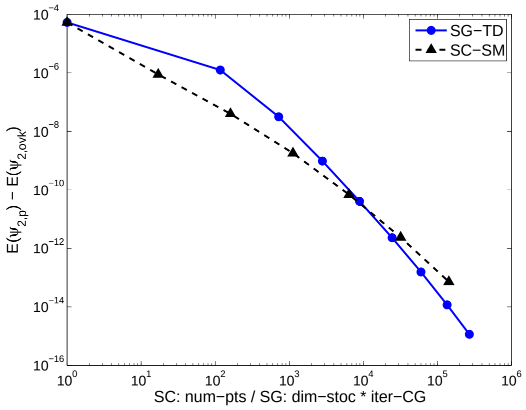


Figure 2.5: Convergence curves for $\varepsilon_{\text{mean}}\,[\psi_2]$ (left) and $\varepsilon_{\text{var}}\,[\psi_2]$ (right) with respect to the computational cost. Comparison between SG-TD and SC-SM methods.

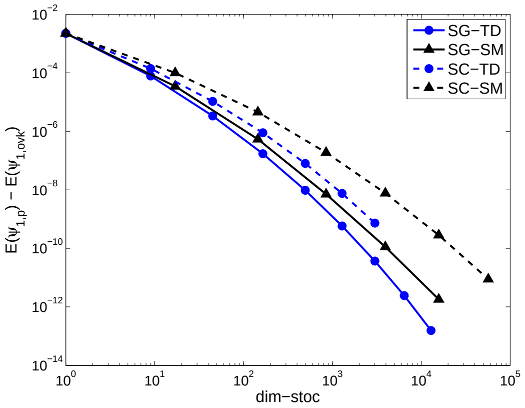


Figure 2.6: Convergence curves for $\varepsilon_{\text{mean}}\,[\psi_1]$ with respect to the dimension of the stochastic space. Comparison between SG and SC methods with TD and SM polynomial spaces.

1; however, since our focus is on anisotropy, we consider only 4 inclusions (the ones in the corners, cf. Fig.2.7-left) so that we can test many different choices of the weights that define the anisotropic spaces (2.10)-(2.13). Nonetheless, the anisotropic setting is particularly meant to be used in high dimensional spaces (see e.g. [72]). For convenience we consider a forcing term uniformly distributed on the whole domain and we look just at $\varepsilon_{\text{mean}}\,[\psi_1]$.

The random coefficient is $a(\mathbf{x},\omega) = 1 + \sum_{n=1}^{4} \gamma_n \chi_n(\mathbf{x}) y_n(\omega)$, with $y_n(\omega) \sim \mathcal{U}(-0.99, 0)$ and $\gamma_n \leq 1$. The values of the coefficients γ_n are shown in Fig. 2.7-left. Notice that these values give different importance to the four random variables. In particular, the inclusion in the bottom-left corner has the largest variance and we expect it to contribute the most to the total variance of the solution. It is therefore intuitively justified to use polynomial degrees higher in the corresponding direction of the stochastic multidimensional space rather than in the other ones. Fig. 2.7 also shows the mean value (middle) and the standard deviation (right) of the solution.

Our goal is to assess the performances of anisotropic polynomial spaces in comparison with their isotropic counterpart. For this we need to estimate the weights to be used in the construction of the anisotropic polynomial space.

We follow closely the argument in [72]. The overall random conductivity coefficient in the n-th inclusion Ω_n is a uniform random variable $\mathcal{U}(a_n, b_n)$ with $a_n = 1 - 0.99\gamma_n$ and $b_n = 1$. This can be

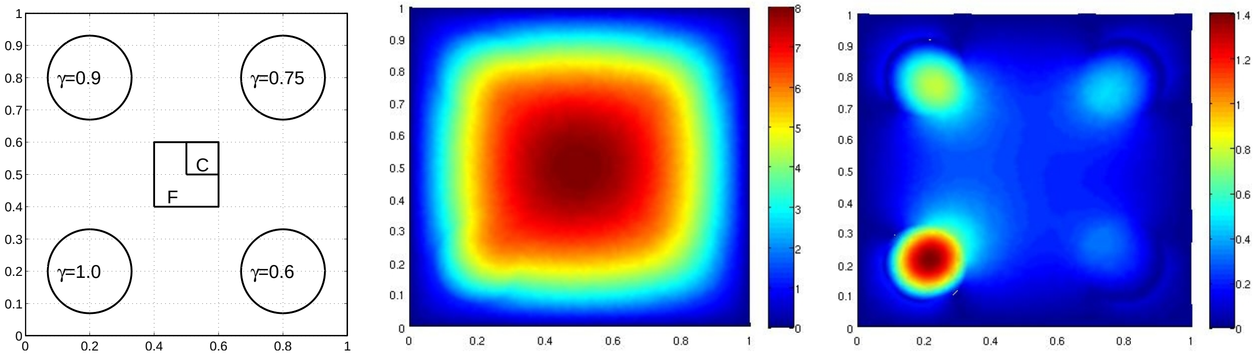


Figure 2.7: Left: geometry for test case 2. Middle: expected value of the solution. Right: standard deviation of the solution.

rewritten as

$$a(\mathbf{x},\omega)_{|\Omega_n} = \frac{a_n+b_n}{2} + \frac{b_n-a_n}{2}\hat{y}_n, \qquad \text{with } \hat{y}_n \sim \mathcal{U}(-1,1).$$

It is easy to show that the solution $u = u(\cdot,\hat{y}_n)$ admits an analytic continuation in the complex region $\Sigma_n = \{z \in \mathbb{C} : \mathfrak{Re}\,(z) > -w_n\}$ with $w_n = \frac{a_n+b_n}{b_n-a_n} = \frac{2-0.99\gamma_n}{0.99\gamma_n}$, which contains, in particular, the interior of the ellipse

$$\mathcal{E}_{\rho_n} = \left\{z \in \mathbb{C} : \mathfrak{Re}\,(z) = \frac{\rho_n+\rho_n^{-1}}{2}\cos\phi,\, \mathfrak{Im}\,(z) = \frac{\rho_n-\rho_n^{-1}}{2}\sin\phi,\, \phi \in [0,2\pi)\right\}$$

with $\rho_n = w_n + \sqrt{w_n^2-1}$.

Standard spectral approximation analysis (see e.g. [25]) allows us to say that interpolation of $u(\cdot,\hat{y}_n)$ in p_n+1 Gauss-Legendre points converges exponentially fast with rate $e^{-g_n p_n}$, with $g_n = \log\rho_n = \log(w_n + \sqrt{w_n^2-1})$.

Therefore the *theoretical estimate* (*a priori* choice) of the weight to be used for the n-th variable is $\alpha_n = g_n$. The larger γ_n, the smaller the corresponding weight α_n. In practice, we have renormalized the weights by dividing them by the smallest one. Notice that the spaces (2.10)-(2.13) remain unchanged by this normalization. The corresponding theoretical weights are in this case $\alpha^{th} = [1,\ 3.5,\ 5.5,\ 7.5]$. To assess the effectiveness of the proposed theoretical estimate, we also consider the weights $\alpha = [1,2,3,4]$ (nearly half the theoretical estimate) and $\alpha = [1,7,11,15]$ (twice the theoretical estimate). Finally, we have also considered an experimental (*a posteriori*) estimate of the coefficients (as suggested in [72]), where the exponential decay $e^{-g_n p_n}$ is estimated numerically by increasing the approximation level in only one direction at a time; the resulting weights are $\alpha^{exp} = [1,\ 2.5,\ 4,\ 5.5]$.

In this example we consider only SG methods in anisotropic TD spaces as they seem to be the most appropriate for this type of problem. Similarly, we restrict our study only to SC methods in the same ATD spaces, so they are directly comparable with the corresponding Galerkin version. The use of SC-ASM methods is expected to give even better results.

We have computed the SG-ATD and SC-ATD with the different choices of weights up to level $w = 21$ and compared them with an overkill solution computed by SG-TD isotropic method at level $w = 22$. This solution has about 14000 sdof . In comparison, the SG-ATD solution has 837 sdof with weights $\alpha = [1,2,3,4]$, 434 sdof with the experimental weights $\alpha^{exp} = [1,2.5,4,5.5]$, 220 sdof with the theoretical weights $\alpha^{th} = [1,3.5,5.5,7.5]$, and 68 sdof with the weights $\alpha = [1,7,11,15]$. We observe that the level $w = 22$ isotropic TD space contains all the ATD spaces with level $w < 22$, therefore our overkill solution is much more accurate than the other ones considered here.

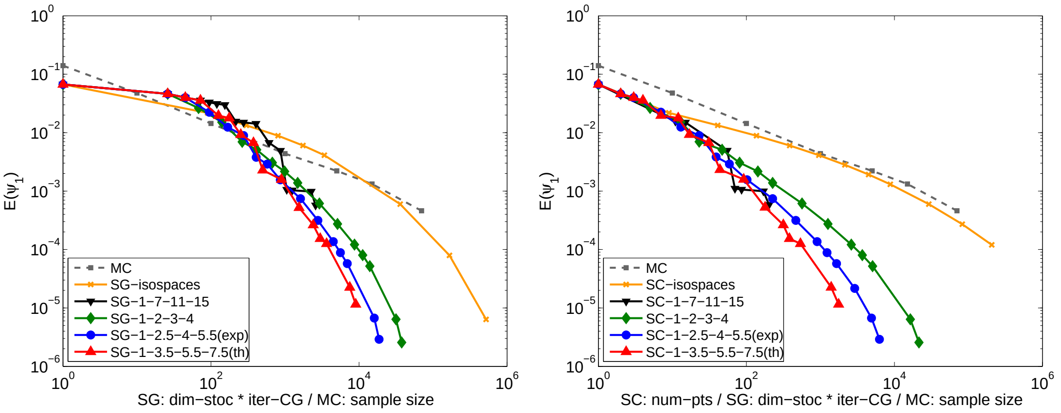


Figure 2.8: Performance of SG-ATD (left) and SC-ATD (right) methods with different choices of weights, in the computation of $\mathbb{E}[\psi_1]$. Error $\varepsilon_{\text{mean}}[\psi_1]$ versus computational cost.

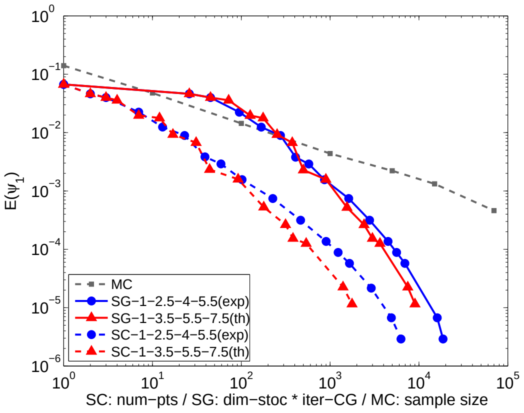


Figure 2.9: Comparison between SG-ATD and SC-ATD methods with best weights in the computation of $\mathbb{E}[\psi_1]$. Error $\varepsilon_{\text{mean}}[\psi_1]$ versus computational cost.

Fig. 2.8 shows the error in computing $\mathbb{E}[\psi_1]$ versus the estimated computational cost when using the SG-ATD (left) or SC-ATD (right) methods. For reference purposes we have also added the convergence plot for Monte Carlo.

First, we observe that SC and SG outperform the standard Monte Carlo. Fig. 2.8 also shows that the theoretical estimate of the weights performs better than all other choices and seems to be very close to optimum for both SC and SG methods, while the *a posteriori* choice gives slightly worse results although the convergence curve is smoother.

In Fig. 2.9 we compare the performances of the SG-ATD and SC-ATD methods with the theoretical and experimental choices of the weights. In this test, the collocation method seems to be superior to the Galerkin one, even for very small tolerances.

2.5 The lognormal case

The goal of this section is to repeat the analysis performed on Test 1 (see section 2.4.a), replacing the uniform random variables with lognormal variables, to see if and how much the probability distribution plays a role in the performance of the Galerkin and Collocation methods. Moreover, the study of the lognormal case is of practical interest, since many hydrological and geological

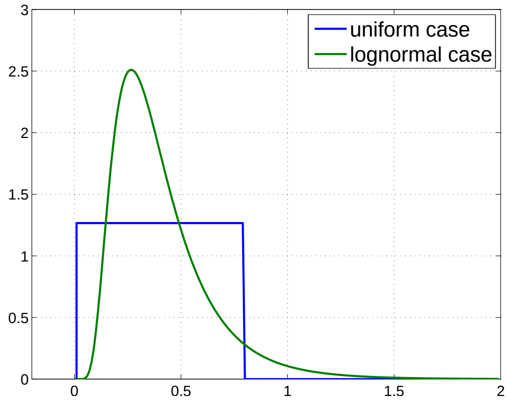


Figure 2.10: comparison of the probability density function for the random inclusions in the uniform case, see eq. (2.23), and lognormal case, see eq. (2.27).

applications typically use such random variables, see (1.10).

A lognormal variable $L(\omega)$ is defined as the exponential of a gaussian random variable: $L = e^{\mu+\sigma Y}$, with $Y \sim \mathcal{N}(0,1)$. It holds $\mathbb{E}\left[L\right] = e^{\mu+\frac{\sigma^2}{2}} = \lambda$ and $\mathbb{V}\mathrm{ar}\left[L\right] = e^{2\mu+\sigma^2}(e^{\sigma^2}-1) = \lambda^2(e^{\sigma^2}-1)$. In this section we will therefore consider the same diffusion coefficient of Section 2.4.a, see eq. (2.23), replacing the uniform variables with lognormal ones,

$$a(\mathbf{x},\omega) = a_0 + \sum_{n=1}^{8} \chi_n(\mathbf{x}) \left(l_0 + e^{\mu+\sigma y_n(\omega)}\right), \qquad y_n \sim \mathcal{N}(0,1) \text{ i.i.d.} \tag{2.27}$$

where now the image set of each random variable $y_n(\omega)$ is $\Gamma_n \equiv \mathbb{R}$ and the parameters a_0, l_0, μ, σ are chosen so that $a(\mathbf{x},\omega)$ is "as close as possible" to its counterpart (2.23). To do this, we set $a_0 = 1$, so that we ensure that the diffusion coefficient outside the inclusions is identical in both cases, and l_0, μ, σ in order to ensure that the inclusions feature the same minimum value, mean and variance as in Test 1; the solution to the corresponding non-linear system can be obtained numerically and is found to be $l_0 = -0.99$, $\mu \approx -1.07$, $\sigma \approx 0.54$, see fig. 2.10.

Note that even if we refer to (2.27) as a lognormal diffusion coefficient, its underlying probability density function $\rho(\mathbf{y}) : \Gamma = \Gamma_1 \times \ldots \times \Gamma_N \to \mathbb{R}$ is indeed a product of Gaussian density functions:

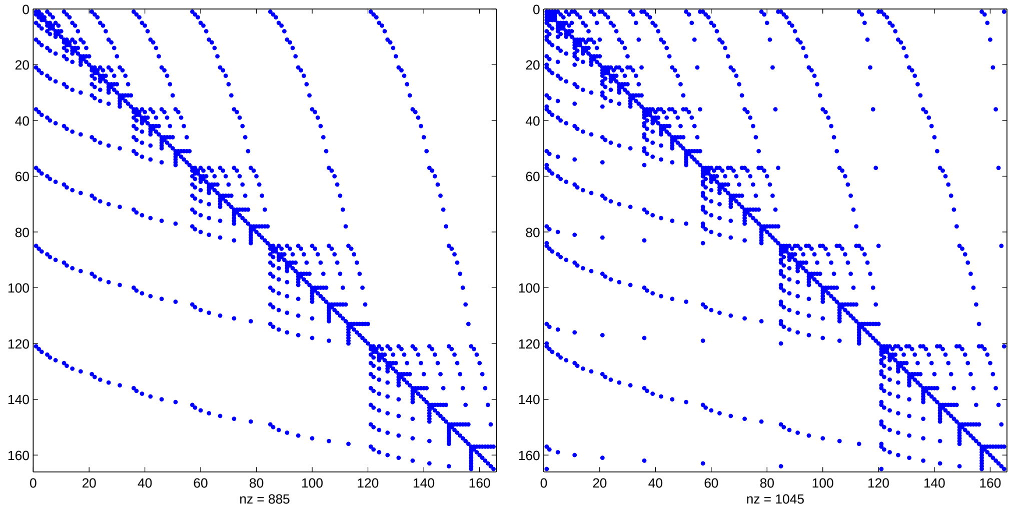
$$\rho(\mathbf{y}) = \prod_{n=1}^{N} \rho_n(y_n), \qquad \rho_n(y_n) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y_n^2}{2}}. \tag{2.28}$$

This is not only much more convenient than expressing $a(\mathbf{x},\omega)$ in terms of a set of lognormal random variables (with this choice we can in fact use the Hermite polynomials for the Galerkin method and Gauss-Hermite quadrature rules for the Collocation method), but is indeed the only possible way of working, as it has been shown in [31] that the sequence of orthonormal polynomials associated to the lognormal density function

$$\rho_L(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{\frac{(\log y-\mu)^2}{2\sigma^2}}$$

is not a complete basis for $L^2_{\rho_L}(\mathbb{R})$. Thus there will be some elements of $L^2_{\rho_L}(\mathbb{R})$ that are not the limit of their expansion in terms of $\rho_L(y)$-orthonormal polynomials.

To set up the Galerkin method, we proceed similarly to Sections 2.3.a. Given $\rho(\mathbf{y})$ as in (2.28), let $\Lambda(w)$ be a sequence of index sets as in (2.6)-(2.13) and $\mathbb{P}_{\Lambda(w)}$ the subspace of $L^2_\rho(\Gamma)$ spanned



(a) uniform case

(b) lognormal case

Figure 2.11: sparsity plots for the SG matrix in the uniform and lognormal cases. In both cases we consider $\Lambda(w)$ =TD(3).

by the set of $\rho(\mathbf{y})d\mathbf{y}$-orthonormal multivariate Hermite polynomials $\{\mathcal{H}_{\mathbf{p}}(\mathbf{y})\}_{\mathbf{p}\in\Lambda(w)}$, that are again computed as products of monodimensional Hermite polynomials $H_p(y)$, see e.g. [39, 85].

The Galerkin method will again require the solution of a block-system like (2.17), where now $G^n_{\mathbf{p},\mathbf{q}}$ is defined as

$$G^n_{\mathbf{p},\mathbf{q}} = \int_\Omega e^{\mu+\sigma y_n}\mathcal{H}_{\mathbf{p}}(\mathbf{y})\mathcal{H}_{\mathbf{q}}(\mathbf{y})\rho(\mathbf{y})d\mathbf{y}.$$

Contrary to (2.24), it is enough for $G^n_{\mathbf{p},\mathbf{q}}$ to be nonzero that $\mathbf{p}$ and $\mathbf{q}$ have all equal components but the $\overline{n}$-th: therefore, the matrix is still block-sparse, but with more nonzero terms than the uniform case, see Figure 2.11 for a comparison of the sparsity patterns in the two cases. The computation of $G^{\overline{n}}_{\mathbf{p},\mathbf{q}}$ can be achieved exactly, since

$$e^\mu \int_{\Gamma_n} e^{\sigma y_n} H_{p_n}(y_n)H_{q_n}(y_n)\frac{1}{\sqrt{2\pi}}e^{-\frac{y_n^2}{2}}dy_n = e^{\mu+\frac{\sigma^2}{2}}\int_{\Gamma_n} H_{p_n}(y_n)H_{q_n}(y_n)\frac{1}{\sqrt{2\pi}}e^{\frac{(y_n-\sigma)^2}{2}}dy_n, \tag{2.29}$$

and the last integral can be computed exactly by a Hermite-Gaussian quadrature rule with $k = (p_n + q_n + 1)/2$ nodes. We have found numerically that many of these terms are small, with larger values of $|p_n - q_n|$ resulting in smaller values of $G^n_{\mathbf{p},\mathbf{q}}$. In this work we have neglected all the terms smaller than 10^{-6} in absolute value.

To solve the Galerkin system we consider again a mean-based preconditioner, see eq. (2.19). However, contrary to the uniform case (2.26), in the lognormal case the preconditioners $P_{\mathbf{q},\mathbf{q}}$ are not all equal, cf. eq. (2.25) and (2.29). Precomputing every Cholesky factorization once and for all before solving the linear system leads to excessive memory requirement, therefore we employ a direct solver at each preconditioning step. Moreover, the mean-based preconditioner is no longer optimal in the case of lognormal random variables and the condition number of the preconditioned matrix grows with the polynomial degree used (see Fig. 2.15). Other preconditioners, specific and more efficient for the lognormal case, have been proposed in [84, 90, 103]. See also [104] for an alternative approach to the numerical treatment for the lognormal case.

The set up of the Stochastic Collocation method is identical to Section 2.3.b; of course we use here quadrature rules based on the roots of Hermite polynomials.

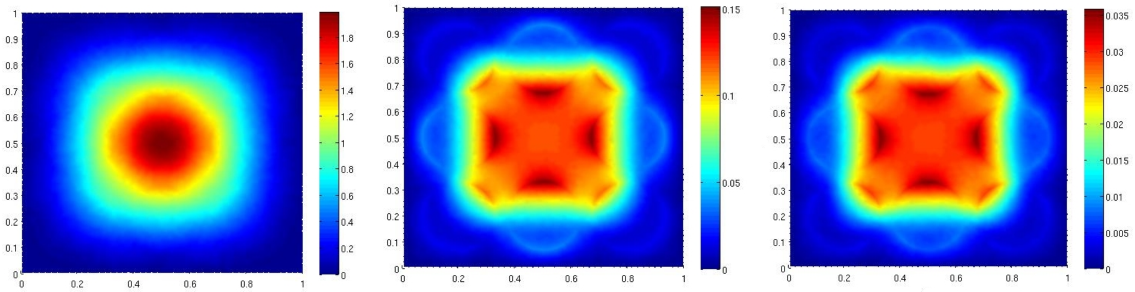


Figure 2.12: expected value (left) and standard deviation (center) of u in the lognormal case. We also show the difference between the standard deviations in the uniform and lognormal cases (right).

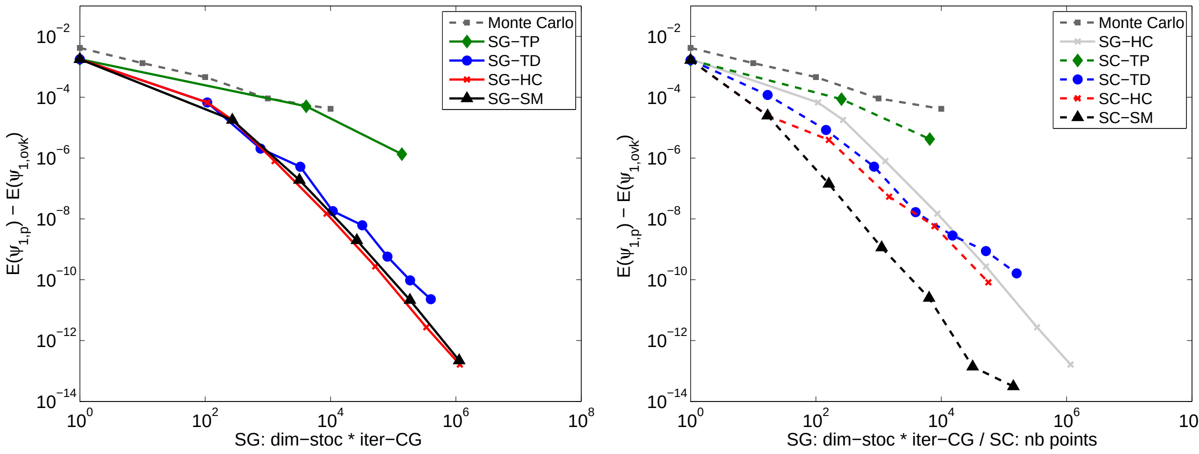


Figure 2.13: Error $\varepsilon_{\text{mean}}[\psi_1]$ versus estimated computational cost for Galerkin (left) and Collocation (right).

Numerical results

The mean and standard deviation field of the solution are qualitatively identical to those in Test 1, cf. Figure 2.2 and 2.12. However, we note that the variance of the solution is slightly lower for the lognormal case (around 20%, see Figure 2.12-right), even if the mean and the variance of the input lognormal random variables have been tuned to have the same mean and variance as the uniform variables of the previous case. We will show in the following how this affects the convergence rate.

We monitor the convergence of the Galerkin and Collocation method in terms of the error measures ε_{mean} applied to the linear functionals ψ_1 and ψ_2, as defined in Section 2.4. Figure 2.13 shows the convergence of $\varepsilon_{\text{mean}}[\psi_1]$ versus estimated computational cost for Galerkin and Collocation. On the one hand, it clearly appears that the TD space in this case is not the best polynomial space where to set the approximation: this suggests that the shape of the probability distribution plays a role in determining the most suitable polynomial space; on the other hand, the SM Collocation scheme is still the most effective one. We also remark that the difference in the performance between Collocation and Galerkin is more evident than in Test 1 (cf. figure 2.3). The same phenomena appear when looking at $\varepsilon_{\text{mean}}[\psi_2]$, see figure 2.14.

The degradation of the performance of the Galerkin is due to the fact that the matrix is less sparse and the mean-based preconditioner is less effective for the lognormal case, see e.g. [84, 90, 103, 104], so that the linear system requires more PCG iterations if compared to the uniform case,

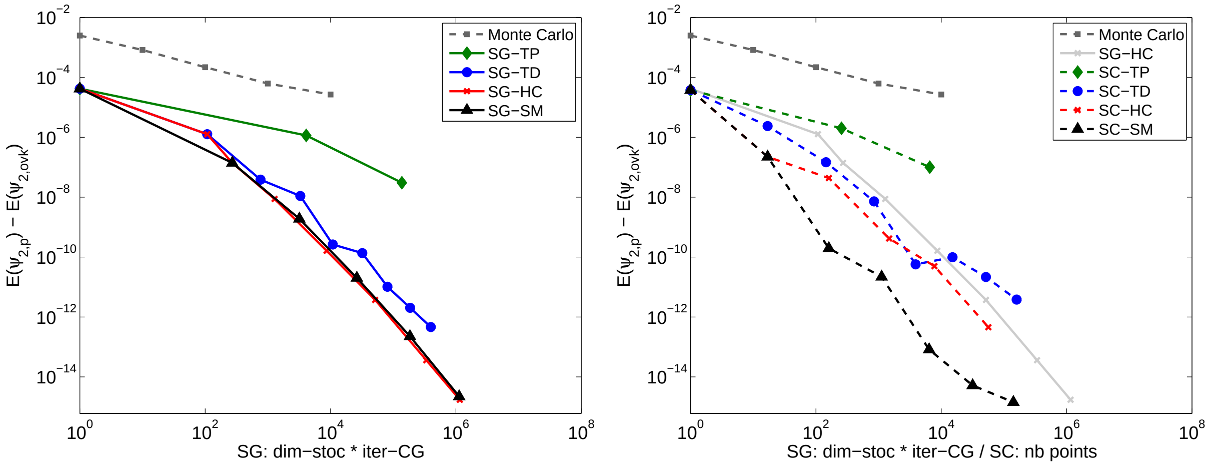


Figure 2.14: Error $\varepsilon_{\text{mean}}[\psi_2]$ versus estimated computational cost for Galerkin (left) and Collocation (right).

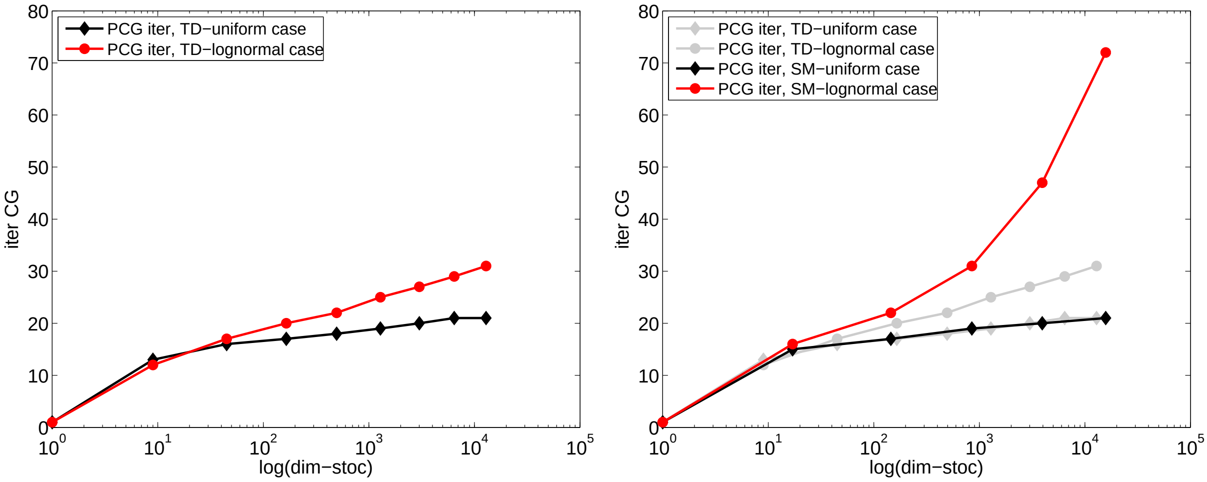
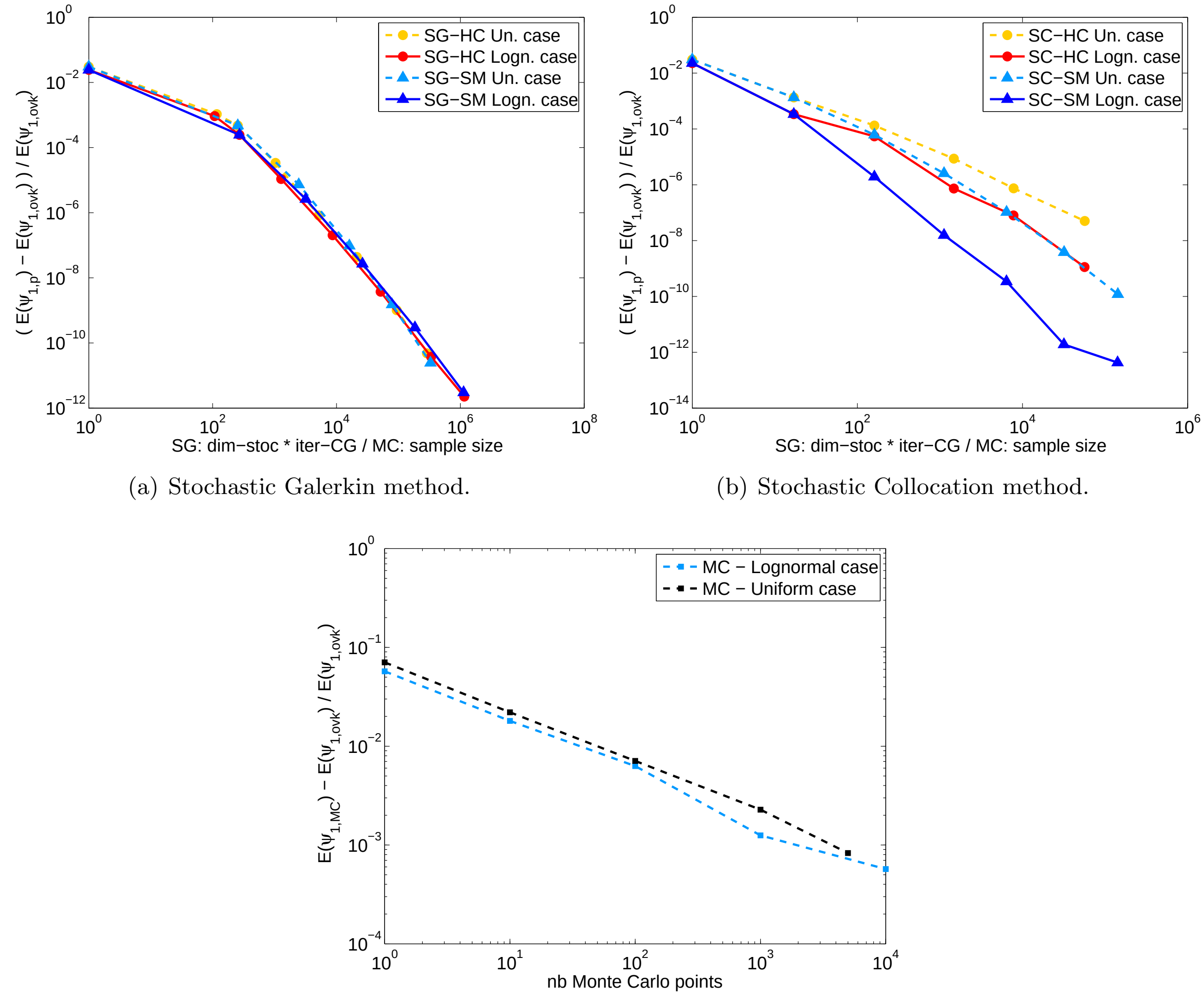


Figure 2.15: Number of PCG iterations of Galerkin method. Left: TD case, right: SM case.

see Figure 2.15. On the other hand, the computational cost of the Collocation is the same for both the tests, since the number of points in the sparse grid is of course the same for both the uniform and the lognormal cases (note that this does not mean that the actual total computational time, which depends on the runtime of each call to the deterministic solver, will be the same in both cases).

We can also compare the convergences of the various methods for the two settings, uniform and lognormal, again in terms of $\varepsilon_{\text{mean}}[\psi_1]$ (normalizing the solutions with respect to the corresponding reference solutions). We thus notice that the comparison gives quite different results depeding on the method. On the one hand, the difference between the standard deviations detected in Figure 2.12 does not appear to influence significantly the convergence of the Monte Carlo method, see Figure 2.16(c), while on the other hand there seems to be a significant improvement of the convergence rates in the Collocation setting, see Figure 2.16(b). The Galerkin method again does not show any difference between the convergence in the two cases, see Figure 2.16(a): this may be due to the compensation of two competing factors: the fact that the polynomial spaces HC, SM may be particularly well suited for the lognormal problem, which positively affects the convergence rate, and



(a) Stochastic Galerkin method.

(b) Stochastic Collocation method.

(c) Monte Carlo method.

Figure 2.16: Comparison of the convergences of the various methods for the uniform and lognormal settings.

the increased computational cost caused by the lower efficiency of the mean-based preconditioner, as pointed out in Figure 2.15.

2.6 Conclusions

In this chapter we have set up an as-fair-as-possible comparison between Collocation and Galerkin method. The effort has been theoretical at first, trying to generalize the classical Smolyak Sparse Grid algorithm for Collocation, so that both the Galerkin and Collocation solutions belong to the same polynomial space: this is a crucial step and represents a novelty in the literature of sparse grids.

Once accomplished this setting, we compared the two methods in terms of accuracy versus computational cost, defined as the number of deterministic problems solved in each case. We compared Galerkin and Collocation considering different random variables in the diffusion coefficients, for several choices of the polynomial spaces, both in isotropic and anisotropic settings. In all cases, the

Collocation method has shown slightly better performances than Galerkin, at least for moderate error tolerances; in particular, the choice of good preconditioners has been found out to be a crucial point in the performances of Galerkin method.

Beside the efficiency in the computation, one also has to compare the differences in the setup costs of the methods. In this sense, Collocation seems to be more appealing since it features similar convergence rates than Galerkin but requires almost no effort moving from a test to another, once the code for sparse grid is available.

Chapter 3

Optimal Galerkin and Collocation approximations

This Chapter consists of the paper by J. Beck, F. Nobile, L. Tamellini, R. Tempone, *On the optimal polynomial approximation of stochastic PDEs by Galerkin and Collocation methods*, to appear on *Mathematical Models and Methods in Applied Sciences*, up to the alignment of the notation and minor improvements in the readibility.

A shorter version with different numerical tests can be found on *ESAIM Proceedings 33(2011), Proceedings of the CANUM conference 2010, Carcans-Maubuisson, France, May 31-June 4, 2010.*

3.1 Introduction

From the previous chapters, it is clear that both the Stochastic Galerkin and Collocation method suffer the so-called "Curse of Dimensionality": using naive projections/interpolations over tensor product polynomials spaces/tensor grids leads to computational costs that grow exponentially fast with the number of random variables. Therefore the main requirement for these methods to be appealing compared to the classical sampling methods (that do not suffer any degradation of the performance when the number of random variables increases) is the capability of retaining good approximations of u while keeping the computational cost as low as possible.

In a Stochastic Galerkin setting this requirement can be translated to the implementation of algorithms able to compute what is known as "best M-terms approximation". In other words, the method should be able to establish a-priori the set of the M most fruitful multivariate orthogonal polynomials in the spectral approximation of u, and to compute only those terms.

Important contributions in the study of the best M-terms approximation have been given by Schwab and co-workers: estimates on the decay of the coefficients of the spectral expansion of u have been proved e.g. in [11, 20, 21]. In this chapter we will reformulate and slightly generalize the result given in [21, Corollary 6.1], and show on few numerical examples that the sequence of polynomial subspaces built upon those estimates ("TD with factorial correction" sets, TD-FC in the following) performs better than classical choices such as Total Degree or Tensor Product in terms of error versus the dimension of the polynomial space.

In a Stochastic Collocation setting, the construction of an optimal grid can be recast to a classical knapsack problem and relies on the notion of profit of each hierarchical surplus composing the sparse grid, as introduced e.g. in [15]. The "Best M-Terms" grid is then the one built with the set of the M most profitable hierarchical surpluses. In this chapter we propose a heuristic a-priori estimate of the profit of each hierarchical surplus, and use it to build a quasi optimal sparse grid. The estimates of the profit are in turn based on the estimates of the decay of the spectral expansion of u. Numerical investigations show that these new grids perform better than standard Smolyak grids as well as grids constructed with the dimension adaptive approach developed in [41, 56]. A similar knapsack approach to the construction of generalized optimal sparse grids has been proposed also in [45]. Our contribution extends and details the procedure to the case of PDEs with stochastic coefficients, working with analytic functions instead of H^r_{mix} ones, and using sharp estimates for the profits of the hierarchical surpluses.

This Chapter is organized as follows. Section 3.2 defines the elliptic model problem of interest and gives general regularity results of the solution u. In Section 3.3 we first address the general procedure that leads to the Stochastic Galerkin approximation of u; next we state the estimate for the decay of the spectral approximation of u and explain how to build practically the TD-FC polynomial subspaces that stem from it. In Section 3.3.b we consider some simple numerical tests where we can build explicitly the best M-terms approximation, and we compare it with the TD-FC and with some standard choices of polynomial subspaces. In Section 3.4 we recall the construction of a general sparse grid, motivate our heuristic estimate of the profit of each hierarchical surplus and explain how to construct in practice optimized sparse grids based on such estimates. Section 3.4.b shows on some simple test cases the effectiveness of the method and the sharpness of our heuristic estimates. Finally Section 3.5 draws some conclusions.

3.2 Problem setting

Let D be a convex bounded polygonal domain in $\mathbb{R}^d$ and $(\Omega, \mathcal{F}, P)$ be a complete probability space. Here Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events and $P : \mathcal{F} \to [0, 1]$ is a probability measure. Consider the stochastic linear elliptic boundary value problem:

Strong Formulation. *find a random function, $u : \overline{D} \times \Omega \to \mathbb{R}$, such that P-almost everywhere in Ω, or in other words almost surely (a.s.), the following equation holds:*

$$\begin{cases} -\operatorname{div}(a(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)) = f(\mathbf{x}) & \mathbf{x} \in D, \\ u(\mathbf{x},\omega) = 0 & \mathbf{x} \in \partial D. \end{cases} \tag{3.1}$$

where the operators div *and* ∇ *imply differentiation with respect to the physical coordinate only.*

We make the following assumptions on the random diffusion coefficient:

Assumption 3.1 (Coercivity and continuity)**.** $a(\mathbf{x}, \omega)$ *is strictly positive and bounded with probability 1, i.e. there exist $a_{min} > 0$ and $a_{max} < \infty$ such that $P(a_{min} \leq a(\mathbf{x}, \omega) \leq a_{max}, \forall \mathbf{x} \in \overline{D}) = 1$.*

Assumption 3.2 (Finite dimensional noise)**.** $a(\mathbf{x}, \omega)$ *is parametrized by a set of N independent and identically distributed uniform random variables in $(-1, 1)$, $\mathbf{y}(\omega) = [y_1(\omega), ..., y_N(\omega)]^T : \Omega \to \mathbb{R}^N$.*

Observe that the assumption that the random variables are uniform is not that restrictive. Indeed, we could assume that a is parametrized by N random variables z_i, $i = 1, \ldots, n$ and introduce a non linear map $y_i = \Theta(z_i)$ that transforms each of them into uniform random variables, following the well known theory on copulas, see e.g. [69].

We denote by $\Gamma_n = (-1, 1)$ the image set of the random variable y_n, and let $\Gamma = \Gamma_1 \times \ldots \times \Gamma_N$. In addition, let $\mathbb{R}_+$ be the set of positive numbers, $\mathbb{R}_+ = \{r \in \mathbb{R} : r > 0\}$, and similarly $\mathbb{R}^N_+ = \{\mathbf{r} \in \mathbb{R}^N : r_i > 0, \forall i = 1, \ldots, N\}$. After Assumption 3.2 the random vector $\mathbf{y}$ has a joint probability density function $\rho : \Gamma \to \mathbb{R}_+$ that factorizes as $\rho(\mathbf{y}) = \prod_{n=1}^N \rho_n(y_n)$, $\forall \mathbf{y} \in \Gamma$, with $\rho_n = \frac{1}{2}$. Moreover, the solution u of (3.1) depends on the single realization $\omega \in \Omega$ only through the value taken by the random vector $\mathbf{y}$. We can therefore replace the probability space $(\Omega, \mathcal{F}, P)$ with $(\Gamma, B(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, where $B(\Gamma)$ denotes the Borel σ-algebra on Γ and $\rho(\mathbf{y})d\mathbf{y}$ is the distribution measure of the vector $\mathbf{y}$. We denote with $L^2_\rho(\Gamma)$ the space of square integrable functions on Γ with respect to the measure $\frac{1}{2^N}d\mathbf{y}$. Note that in the case the original random variables are not uniform but with bounded support, and a mapping Θ is not available, one could still reduce the problem to the uniform case, at the price of bounding $\rho(\mathbf{y})$ with $\|\rho(\mathbf{y})\|_\infty$.

The assumption of independence of the random variables is very convenient for the development of the techniques proposed below, since they rely on tensor polynomial approximations. However, such assumption is not essential and could be removed whenever the density ρ does not factorize, by introducing an auxiliary density $\hat{\rho} = \frac{1}{2^N}$ as suggested in [4]. The price to pay in the convergence estimate is then a costant factor proportional to $\|\rho/\hat{\rho}\|_{L^\infty(\Omega)}$.

In the rest of the chapter we will use the following notation: given a multi-index $\mathbf{i} \in \mathbb{N}^N$ and a vector $\mathbf{r} \in \mathbb{R}^N$, we define $|\mathbf{i}| = \sum_{n=1}^N i_n$, $\mathbf{i}! = \prod_{n=1}^N (i_n!)$ and $\mathbf{r}^\mathbf{i} = \prod_{n=1}^N r_n^{i_n}$. We can now state a regularity assumption on $a(\mathbf{x}, \mathbf{y})$:

Assumption 3.3 (Stochastic regularity). *$a(\mathbf{x}, \mathbf{y})$ is infinitely many times differentiable with respect to $\mathbf{y}$ and $\exists\, \mathbf{r} \in \mathbb{R}^N_+$ s.t.*

$$\left\| \frac{\partial_\mathbf{i} a}{a}(\cdot, \mathbf{y}) \right\|_{L^\infty(D)} \leq \mathbf{r}^\mathbf{i} \quad \forall \mathbf{y} \in \Gamma,$$

where $\mathbf{i}$ is a multi-index in $\mathbb{N}^N$, $\partial_\mathbf{i} a = \dfrac{\partial^{i_1+\ldots+i_N} a}{\partial y_1^{i_1} \cdots \partial y_N^{i_N}}$, and $\mathbf{r}$ is independent of $\mathbf{y}$.

Example 3.1 (Stochastic regularity). *A common situation of interest is when $a(\mathbf{x}, \omega)$ is an infinitely dimensional random field, suitably expanded in series (e.g. by a Karhunen-Loève or Fourier expansion) either as a* linear *expansion of the form $a = a_0 + \sum_{n=1}^\infty b_n(\mathbf{x}) y_n$ with $b_n \in L^\infty(D)$ and $a_{min} = a_0 - \sum_{n=1}^\infty \|b_n\|_{L^\infty(D)}$, or an* exponential *expansion of the form $a = a_0 + \exp\left(\sum_{n=1}^\infty b_n(\mathbf{x}) y_n\right)$. Then the infinite series is truncated up to N terms, with N large enough to take into account a sufficiently large amount of the total variability. Both expansions comply with Assumption 3.3 taking $r_n = \|b_n\|_{L^\infty(D)}/a_{min}$ and $r_n = \|b_n\|_{L^\infty(D)}$, respectively.*

Finally, we denote by $V = H^1_0(D)$ the space of square integrable functions in D with square integrable distributional derivatives and with zero trace on the boundary, equipped with the gradient norm $\|v\|_V = \|\nabla v\|_{L^2(D)}$, $\forall v \in V$. Its dual space will be denoted by V'. We are now in the position to write a weak formulation of problem (3.1):

Weak Formulation. *Find $u \in V \otimes L^2_\rho(\Gamma)$ such that $\forall\, v \in V \otimes L^2_\rho(\Gamma)$*

$$\int_\Gamma \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \tag{3.2}$$

Under Assumption 3.1, the Lax-Milgram lemma yields that there exists a unique solution to problem (3.2) for any $f \in V'$. Moreover, the following estimate holds:

$$\|u\|_{V \otimes L^2_\rho(\Gamma)} \leq \frac{\|f\|_{V'}}{a_{min}}.$$

The solution u can also be thought as a function defined in Γ with solution in V, $u : \Gamma \to V$ and, thanks to the previous result, we have $u \in L^2_\rho(\Gamma, V) = V \otimes L^2_\rho$. In what follows we will often use the notation $u(\mathbf{y}) := u(\cdot, \mathbf{y}) \in V$ if no confusion arises.

Concerning the regularity of the solution with respect to $\mathbf{y}$, the following result holds, which generalizes the result given in [21] for the special case $a = a_0 + \sum_{n=1}^N b_n(x) y_n$.

Theorem 3.1. *Let $a(\mathbf{x}, \mathbf{y})$ be a diffusion coefficient for equation* (3.1) *that satisfies Assumptions 3.1 - 3.3. Then the derivatives of u can be bounded as*

$$\|\partial_{\mathbf{i}} u(\mathbf{y})\|_V \leq C_0 |\mathbf{i}|! \, \tilde{\mathbf{r}}^{\mathbf{i}} \quad \forall \mathbf{y} \in \Gamma.$$

Here $C_0 = \dfrac{\|f\|_{V'}}{a_{min}}$ *and* $\tilde{\mathbf{r}} = \left(\dfrac{1}{\log 2}\right) \mathbf{r}$*, with* $\mathbf{r}$ *as in Assumption 3.3.*

The proof is technical; we thus postpone it to the Appendix. A consequence of Theorem 3.1 is that u is analytic in every $\mathbf{y} \in \Gamma$. Of course, since u is understood here as a V-valued function, the notion of analyticity has to be intended accordingly:

Definition 3.1. *A function $f : D \subseteq \mathbb{R}^N \to V$ is said to be analytic if for every $\mathbf{y}_0 \in D$ the Taylor expansion of f centered in $\mathbf{y}_0$ converges to $f(\mathbf{y})$ in V-sense in a neighborhood of $\mathbf{y}_0$.*

Corollary 3.1. *Under the hypotheses of Theorem 3.1, given $\varepsilon > 0$, for every $\mathbf{y}_0 \in \Gamma$ the Taylor series of u converges in the disk*

$$\mathcal{D}(\mathbf{y}_0) = \left\{ \mathbf{y} \in \mathbb{R}^N : \tilde{\mathbf{r}} \cdot \mathrm{abs}\,(\mathbf{y} - \mathbf{y}_0) < 1 \right\}.$$

where $\mathrm{abs}\,(\mathbf{v}) = (|v_1|, \ldots, |v_N|)^T$. *Therefore $u : \Gamma \to V$ is analytic and can be extended analytically to the set*

$$\Sigma = \left\{ \mathbf{y} \in \mathbb{R}^N : \exists \, \mathbf{y}_0 \in \Gamma \, \text{s.t.} \; \tilde{\mathbf{r}} \cdot \mathrm{abs}\,(\mathbf{y} - \mathbf{y}_0) < 1 \right\}.$$

Proof. It is enough to bound the series of V-norms. Use first Theorem 3.1 to bound the norm of the Taylor expansion of $u(\mathbf{y})$ centered in $\mathbf{y}_0 \in \Gamma$ as

$$\left\| \sum_{k=0}^{\infty} \sum_{|\mathbf{j}|=k} \frac{\partial_{\mathbf{j}} u(\mathbf{y}_0)}{\mathbf{j}!} (\mathbf{y} - \mathbf{y}_0)^{\mathbf{j}} \right\|_V \leq \sum_{k=0}^{\infty} \sum_{|\mathbf{j}|=k} C_0 \tilde{\mathbf{r}}^{\mathbf{j}} \frac{|\mathbf{j}|!}{\mathbf{j}!} \, \mathrm{abs}\,(\mathbf{y} - \mathbf{y}_0)^{\mathbf{j}}.$$

Next exploit the generalized Newton binomial formula, that states that for $\alpha_1, \ldots, \alpha_N \in \mathbb{R}_+$ and $k \in \mathbb{N}$ we have

$$\sum_{|\mathbf{j}|=k} \frac{k!}{\mathbf{j}!} \boldsymbol{\alpha}^{\mathbf{j}} = \left(\sum_{n=1}^{N} \alpha_n \right)^k,$$

to rewrite the bound on the norm of the Taylor series as

$$\left\| \sum_{k=0}^{\infty} \sum_{|\mathbf{j}|=k} \frac{\partial_{\mathbf{j}} u(\mathbf{y}_0)}{\mathbf{j}!} (\mathbf{y} - \mathbf{y}_0)^{\mathbf{j}} \right\|_V \leq C_0 \sum_{k=0}^{\infty} \left(\sum_{n=1}^{N} \tilde{r}_n |y_n - y_{0,n}| \right)^k.$$

Thus the Taylor series of u converges to u in the disk $\mathcal{D}(\mathbf{y}_0)$. Therefore u is analytic and admits an analytic extension in Σ. □

3.3 Stochastic Galerkin method

We now seek an approximation of the solution u with respect to $\mathbf{y}$ by global polynomials.

As anticipated in the introduction, we remark that the choice of the polynomial space is critical when the number N of input random variables is large, since the number of stochastic degrees of freedom might grow very quickly with N, even exponentially when isotropic tensor product polynomial spaces are used (see Table 3.1). This effect is known as the *curse of dimensionality.*

Several choices of polynomial spaces that mitigate this phenomenon have been proposed in the literature, see e.g. [5]. Each of these polynomial spaces is built as the span of a properly selected subset of a multivariate orthonormal polynomial basis $\{\mathcal{L}_p(\mathbf{y})\}_{p\in\mathbb{N}}$ for $L^2_\rho(\Gamma)$, to retain good approximating properties with only a finite number of basis functions.

Since $L^2_\rho(\Gamma) = \bigotimes_{n=1}^N L^2_{\rho_n}(\Gamma_n)$, the elements of an orthonormal basis can be built as products of orthonormal polynomials for each of the directions y_n, $\{L_{p_n}(y_n)\}_{p_n\in\mathbb{N}}$; we can thus index the multivariate orthonormal polynomial basis functions $\mathcal{L}_p(\mathbf{y})$ with multi-indices $\mathbf{p} = (p_1, \dots, p_N)$

$$\mathcal{L}_{\mathbf{p}}(\mathbf{y}) = \prod_{n=1}^N L_{p_n}(y_n).$$

Then, by construction, the set $\{\mathcal{L}_{\mathbf{p}}(\mathbf{y})\}_{\mathbf{p}\in\mathbb{N}^N}$ is a $\rho(\mathbf{y})d\mathbf{y}$-orthonormal basis in $L^2_\rho(\Gamma)$, i.e. such that $\int_\Gamma \mathcal{L}_{\mathbf{p}}(\mathbf{y})\mathcal{L}_{\mathbf{q}}(\mathbf{y})\rho(\mathbf{y})\, d\mathbf{y} = 1$ if $\mathbf{p} = \mathbf{q}$ and 0 otherwise.

Let now $w \in \mathbb{N}$ be an integer index indicating the level of approximation, and $\Lambda(w)$ a sequence of increasing index sets such that

$$\Lambda(0) = \{(0, \dots, 0)\}, \quad \Lambda(w) \subseteq \Lambda(w+1) \subset \mathbb{N}^N \text{ for } w \geq 0 \quad \text{and } \mathbb{N}^N = \bigcup_{w\in\mathbb{N}} \Lambda(w). \tag{3.3}$$

Denoting by $\mathbb{P}_{\Lambda(w)}(\Gamma)$ the multivariate polynomial space

$$\mathbb{P}_{\Lambda(w)}(\Gamma) = \mathit{span}\,\{\mathcal{L}_{\mathbf{p}}(\mathbf{y}),\ \mathbf{p} \in \Lambda(w)\}\,, \tag{3.4}$$

the Stochastic Galerkin (SG) approximation consists in restricting the weak formulation (3.2) to the subspace $V \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$ and reads:

Galerkin Formulation. *Find* $u_w \in V \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$ *such that* $\forall\, v_w \in V \otimes \mathbb{P}_{\Lambda(w)}(\Gamma)$

$$\int_\Gamma \int_D a(\mathbf{x}, \mathbf{y})\nabla u_w(\mathbf{x}, \mathbf{y}) \cdot \nabla v_w(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v_w(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}, \tag{3.5}$$

where, due to the orthonormality of $\{\mathcal{L}_{\mathbf{p}}(\mathbf{y})\}_{\mathbf{p}\in\Lambda(w)}$,

$$u_w(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{p}\in\Lambda(w)} u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y}), \quad \text{with} \quad u_{\mathbf{p}}(\mathbf{x}) = \int_\Gamma u(\mathbf{x}, \mathbf{y})\mathcal{L}_{\mathbf{p}}(\mathbf{y})\rho(\mathbf{y})d\mathbf{y} \quad \forall \mathbf{p} \in \Lambda(w). \tag{3.6}$$

Commonly used spaces $\mathbb{P}_{\Lambda(w)}(\Gamma)$ are listed in Table 3.1; for further details, see [5] and references therein.

	index set $\Lambda(w)$	Dimension $\lvert\Lambda(w)\rvert$
Tensor product	$\{\mathbf{p} \in \mathbb{N}^N : \ \max_{n=1\dots,N} p_n \leq w\}$	$(1+w)^N$
Total degree	$\left\{\mathbf{p} \in \mathbb{N}^N : \ \sum_{n=1}^N p_n \leq w\right\}$	$\binom{N+w}{N}$
Hyperbolic cross	$\left\{\mathbf{p} \in \mathbb{N}^N : \ \prod_{n=1}^N (p_n+1) \leq w+1\right\}$	$(w+1)(1+\log(w+1))^{N-1}$

Table 3.1: Examples of typical polynomial spaces. The result for HC is only an upper bound.

One could also consider anisotropic versions of these spaces (see e.g. [3, 5, 72]) as in Table 3.2, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N_+$ is a vector of positive weights and $\alpha_{min} = \min_n \alpha_n$. We can

Tensor product	$\Lambda(w) = \{\mathbf{p} \in \mathbb{N}^N : \ \max_{n=1...,N} \alpha_n p_n \leq \alpha_{min} w\}$
Total degree	$\Lambda(w) = \left\{\mathbf{p} \in \mathbb{N}^N : \ \sum_{n=1}^N \alpha_n p_n \leq \alpha_{min} w\right\}$
Hyperbolic cross	$\Lambda(w) = \left\{\mathbf{p} \in \mathbb{N}^N : \ \prod_{n=1}^N (p_n + 1)^{\frac{\alpha_n}{\alpha_{min}}} \leq w + 1\right\}$

Table 3.2: Corresponding anisotropic version of the polynomial spaces on Table 3.1.

interpret these weights as a measure of the importance of each random variable y_n on the solution: the smaller the weight, the higher degree we allow in the corresponding variable.

The family of orthonormal monodimensional polynomials will of course depend on the measure of each Γ_n (*Generalized Polynomial Chaos*). In the case of uniform random variables, one can use the well-known orthonormal Legendre polynomials; the p-th Legendre polynomial can be computed recursively (see e.g. [39]), or explicitly with the Rodrigues' formula:

$$L_{p_n}(t) = \frac{(-1)^n \sqrt{2p_n + 1}}{2^{p_n} p_n!} \frac{d^{p_n}}{dt^{p_n}} \left((1 - t^2)^{p_n}\right). \tag{3.7}$$

We recall Hermite polynomials for Gaussian measures and Laguerre polynomials for Exponential measures; see [111] for the general Askey scheme. Necessary conditions for the convergence of the Generalized Polynomial Chaos expansion can be found e.g. in [31].

Now let $\phi(\mathbf{x})$ be a basis function for the physical space V. Inserting $v_w = \phi(\mathbf{x})\mathcal{L}_\mathbf{q}(\mathbf{y})$ with $\mathbf{q} \in \Lambda(w)$ as test functions in the weak formulation (3.5) will result in a set of equations in weak form for the coefficients $u_\mathbf{p}(\mathbf{x})$ that will be generally coupled due to the term $a(\mathbf{x}, \mathbf{y})\mathcal{L}_\mathbf{p}(\mathbf{y})\mathcal{L}_\mathbf{q}(\mathbf{y})$ in the equation (3.5). See for instance the works [5, 82, 83] for further details on space discretization and on the numerical solution of such system of equations.

3.3.a Quasi-optimal choice of polynomial spaces

A question that naturally arises in the context of Galerkin approximation concerns the best choice of the polynomial space to be used, to get maximum accuracy for a given dimension M of the space (*best M-terms* approximation). In other words, we look for an index set $\mathcal{S}_M \subset \mathbb{N}^N$ with cardinality M that minimizes the projection error

$$\|u - \sum_{\mathbf{p} \in \mathcal{S}_M} u_\mathbf{p} \mathcal{L}_\mathbf{p}\|^2_{V \otimes L^2_\rho(\Gamma)} = \sum_{\mathbf{p} \notin \mathcal{S}_M} \|u_\mathbf{p}\|^2_V, \tag{3.8}$$

where the equivalence is a consequence of Parseval's equality and the completeness of $\{\mathcal{L}_\mathbf{p}\}_{\mathbf{p} \in \Lambda(w)}$ in $L^2_\rho(\Gamma)$.

Abstract construction

The obvious solution to this problem is to take the set $\mathcal{S}_M$ that contains the M coefficients $u_\mathbf{p}$ with largest norm. This solution of course is not constructive; what we need are sharp estimates of the decay of the coefficients $\|u_\mathbf{p}\|_V$, based only on computable quantities, to be used in the approximation of the set $\mathcal{S}_M$. Actually, assuming that an estimate of the type

$$\|u_\mathbf{p}\|_V \leq G(\mathbf{p}) \tag{3.9}$$

is available, one can define an index set Λ_ϵ by selecting all multi-indices $\mathbf{p}$ for which the *estimated decay* of the corresponding Legendre coefficient is above a fixed threshold $\epsilon \in \mathbb{R}_+$,

$$\Lambda_\epsilon = \left\{\mathbf{p} \in \mathbb{N}^N : G(\mathbf{p}) \geq \epsilon\right\},$$

or equivalently

$$\Lambda(w) = \left\{ \mathbf{p} \in \mathbb{N}^N : -\log G(\mathbf{p}) \leq w,\, w = \lceil -\log \epsilon \rceil \right\}. \tag{3.10}$$

If the sequence $\Lambda(w)$ covers $\mathbb{N}^N$ as w goes to infinity, the corresponding u_w will converge to u and, if the bound $G(\mathbf{p})$ in (3.9) is sharp, $\Lambda(w)$ will be a "quasi optimal" approximation of the best M-terms approximation, where now M denotes the cardinality of $\Lambda(w)$.

A preliminary example

Assume for a moment that u factorizes, i.e. it can be written as a product of 1D analytic functions in the stochastic variables, $u(\mathbf{x},\mathbf{y}) = f(\mathbf{x}) \prod_{n=1}^N v_n(y_n)$. If we denote with v_{n,p_n} the Legendre coefficients of the factor v_n, i.e.

$$v_{n,p_n} = \int_{\Gamma_n} v_n(y_n) L_{p_n}(y_n) \rho_n(y_n) dy_n,$$

the Legendre coefficients of u are given simply by

$$u_{\mathbf{p}}(\mathbf{x}) = f(\mathbf{x}) \prod_{n=1}^N v_{n,p_n}. \tag{3.11}$$

Now, from classical approximation theory ([25, 100]) it is well known that, if v_n is analytic in Γ_n, the coefficient v_{n,p_n} is exponentially decaying in p_n with a certain rate g_n, $|v_{n,p_n}| \leq c(g_n) e^{-g_n p_n}$; as a consequence we easily obtain a sharp bound on the Legendre coefficients of u,

$$\|u_{\mathbf{p}}\|_V \leq \|f\|_V \, \mathrm{C}\, e^{-\sum_n g_n p_n}, \quad \mathrm{C} = \prod_{n=1}^N c\,(g_n). \tag{3.12}$$

Substituting this bound in (3.10), we get that a quasi optimal choice of polynomial sets for a separable function of the form (3.11) is the anisotropic TD sets sequence defined in Table 3.2 with weights $\alpha_n = g_n$.

General case

In the general case deriving sharp estimates on the decay of $\|u_{\mathbf{p}}\|_V$ is a more delicate task. Seminal works in this direction are [11, 21, 20], where estimates of the decay of the Legendre coefficients are provided. We consider here a slight generalization of the result in [21, Corollary 6.1] and show numerically that the polynomial sets built on these modified estimates behave closely to the true best M-terms approximation.

Under Assumptions 3.1 - 3.3 it is possible to prove that the following estimate holds for the Legendre coefficients. Again, a similar result is given in [21] for the special case $a = a_0 + \sum_{n=1}^N b_n(x) y_n$.

Proposition 3.1. *Consider equation* (3.1)*, suppose that the diffusion coefficient a satisfies Assumptions 3.1 - 3.3, let $\mathbf{r}$ be as in Assumption 3.3 and C_0 be as in Theorem 3.1. Then the V-norm of the Legendre coefficients $u_{\mathbf{p}}$ can be bounded as*

$$\|u_{\mathbf{p}}\|_V \leq C_0 e^{-\sum_n g_n p_n} \frac{|\mathbf{p}|!}{\mathbf{p}!}, \quad g_n = -\log(\, r_n/(\sqrt{3}\log 2)\,). \tag{3.13}$$

Proof. We follow closely the proof in [21, Corollary 6.1]. We start from the definition of the Legendre coefficients (3.6) and the Rodrigues' formula for the Legendre polynomials (3.7). Integrating by parts and thanks to the properties of the Bochner integral we have

$$\begin{aligned}\|u_\mathbf{p}\|_{V(D)} &= \left\| \int_\Gamma u(\cdot,\mathbf{y})\mathcal{L}_\mathbf{p}(\mathbf{y})\rho(\mathbf{y})d\mathbf{y} \right\|_V \\ &\le \frac{\prod_{n=1}^N \sqrt{2p_n+1}}{2^{|\mathbf{p}|}\mathbf{p}!} \int_\Gamma \left\|\partial_\mathbf{p}^\mathbf{y} u(\cdot,\mathbf{y})\right\|_V \prod_{n=1}^N (1-y_n^2)^{p_n}\rho(\mathbf{y})d\mathbf{y}.\end{aligned}$$

It has been shown in [21] that

$$I(\mathbf{p}) = \prod_{n=1}^N \sqrt{2p_n+1} \int_\Gamma \prod_{n=1}^N (1-y_n^2)^{p_n}\rho(\mathbf{y})d\mathbf{y} \le \left(\frac{2}{\sqrt{3}}\right)^{|\mathbf{p}|}. \tag{3.14}$$

Thus we have

$$\|u_\mathbf{p}\|_{V(D)} \le \max_{\mathbf{y}\in\Gamma} \left\|\partial_\mathbf{p}^\mathbf{y} u(\cdot,\mathbf{y})\right\|_V I(\mathbf{p})\frac{1}{2^{|\mathbf{p}|}\mathbf{p}!},$$

and the proof is completed using Theorem 3.1 to estimate $\max_{\mathbf{y}\in\Gamma}\left\|\partial_\mathbf{p}^\mathbf{y} u(\cdot,\mathbf{y})\right\|_V$:

$$\begin{aligned}\|u_\mathbf{p}\|_{V(D)} &\le C_0|\mathbf{p}|! \left(\frac{1}{\log 2}\mathbf{r}\right)^\mathbf{p} \left(\frac{2}{\sqrt{3}}\right)^{|\mathbf{p}|} \frac{1}{2^{|\mathbf{p}|}\mathbf{p}!} \\ &= C_0 \left(\frac{1}{\sqrt{3}\log 2}\mathbf{r}\right)^\mathbf{p} \frac{|\mathbf{p}|!}{\mathbf{p}!} = C_0 e^{\sum_n p_n \log\left(\frac{r_n}{\sqrt{3}\log 2}\right)} \frac{|\mathbf{p}|!}{\mathbf{p}!}.\end{aligned} \tag{3.15}$$

□

Example 3.2. *To motivate bound* (3.13)*, assume that in the model problem* (3.1) *the forcing term is deterministic, $f = f(\mathbf{x})$, and the diffusion coefficient is constant in space, $a = a(\mathbf{y}) = 1 + \sum_{i=1}^N b_i y_i$, with $b_i > 0$. As explained in Remark 3.1, for such a diffusion coefficient Assumption 3.3 holds with $a_{min} = 1 - \sum_i b_i$ and $r_i = b_i/a_{min}$. Moreover, let us denote with $g \in V$ the solution of the auxiliary problem*

$$\begin{cases} \Delta g(\mathbf{x}) = f(\mathbf{x}) & \mathbf{x} \in D, \\ g(\mathbf{x}) = 0 & \mathbf{x} \in \partial D. \end{cases}$$

Under these hypotheses we can derive an analytic expression for u and its derivatives with respect to $\mathbf{y}$,

$$u(\mathbf{x},\mathbf{y}) = g(\mathbf{x})\frac{1}{1+\sum_{i=1}^N b_i y_i}, \quad \partial_\mathbf{p} u(\mathbf{x},\mathbf{y}) = g(\mathbf{x})\frac{|\mathbf{p}|!\ \mathbf{b}^\mathbf{p}}{\left(1+\sum_{i=1}^N b_i y_i\right)^{|\mathbf{p}|+1}}. \tag{3.16}$$

We can exploit this fact to compute explicitly a bound for the V-norm of the Legendre coefficients $u_\mathbf{p}$ of u. Actually, using again Rodrigues' formula (3.7) *in the definition of $u_\mathbf{p}$, and integrating by parts, we obtain*

$$u_\mathbf{p}(\mathbf{x}) = \int_\Gamma u(\mathbf{x},\mathbf{y})\mathcal{L}_\mathbf{p}(\mathbf{y})\rho(\mathbf{y})d\mathbf{y} \le g(\mathbf{x})\frac{\mathbf{b}^\mathbf{p}}{2^{|\mathbf{p}|}}\frac{|\mathbf{p}|!}{\mathbf{p}!}\frac{1}{(a_{min})^{|\mathbf{p}|+1}}\prod_{i=1}^N (-1)^n \sqrt{2p_n+1}\int_{[-1,1]} (1-y_n^2)^{p_n}\frac{1}{2}dy_n.$$

Finally we exploit bound (3.14)*, pass to the V-norm and use the fact that $\|g\|_V = \|f\|_{V'}$ to obtain*

$$\|u_\mathbf{p}\|_V \le \frac{\|f\|_{V'}}{a_{min}}\frac{\mathbf{b}^\mathbf{p}}{(a_{min})^{|\mathbf{p}|}}\left(\frac{1}{\sqrt{3}}\right)^{|\mathbf{p}|}\frac{|\mathbf{p}|!}{\mathbf{p}!}.$$

This can be recast using the definition of the rate $r_i = b_i/a_{min}$ and of the constant C_0 in Theorem 3.1 to

$$\|u_{\mathbf{p}}\|_V \leq C_0 e^{-\sum_n g_n p_n} \frac{|\mathbf{p}|!}{\mathbf{p}!}, \quad g_n = -\log(r_n/\sqrt{3}),$$

that is precisely the bound derived in Proposition 3.1, with a slight modification on the rate g_i. Numerical results in the next Section will again cover this particular example, showing that the bound proposed yields good approximating properties.

Remark 3.1. *Since $u(\mathbf{x}, \cdot)$ is analytic in Γ (see Corollary 3.1), it can be shown that u always admits a converging Legendre expansion. In spite of this, the estimate* (3.13) *in the previous Proposition does not ensure that the norm of the coefficients $\|u_{\mathbf{p}}\|_V$ of the expansion is decaying for any value of the coefficients r_n when $|\mathbf{p}| \to \infty$, nor that the Legendre series is convergent; sufficient conditions for this to be true are given in the next Preposition.*

This is a clear indication that estimate (3.13) *is not sharp. Other estimates derived using complex analysis arguments are available and always predict a decay of $\|u_{\mathbf{p}}\|_V$ for $|\mathbf{p}| \to \infty$ (see e.g. [20]). On the other hand, we have observed that the behaviour of the Legendre coefficients is well described by a bound of the type of* (3.13)*, if the rates g_n are estimated numerically rather than analytically. See Section 3.3.b for numerical evidence on the quality of the bound proposed.*

For a given set Λ, let $\overline{w}_\Lambda$ be the index of the largest TD set included in Λ:

$$\overline{w}_\Lambda = \max\{\widetilde{w} \in \mathbb{N} : TD(\widetilde{w}) \subseteq \Lambda(w)\}. \tag{3.17}$$

The following Proposition holds:

Proposition 3.2. *Given an increasing sequence of index sets $\Lambda(w)$ with $\overline{w} \to \infty$, the estimate* (3.13) *in Proposition 3.1 implies that a sufficient condition for the Legendre series u_w defined in* (3.6) *to converge uniformly to u is*

$$\sum_{i=1}^{N} r_n < \log 2\,. \tag{3.18}$$

Proof. It is enough to prove that if condition (3.18) holds then the sequence $u_w = \sum_{\mathbf{p}\in\Lambda(w)} u_{\mathbf{p}}\mathcal{L}_{\mathbf{p}}$ is Cauchy with respect to the norm $\|\cdot\|_{L^\infty(\Gamma;V)}$, for the sequence $\Lambda(w)$ considered. As a consequence u_w converges uniformly to its limit u.

To prove that u_w is Cauchy, let $w_1, w_2 \in \mathbb{N}$ such that $w_1 < w_2$. Moreover, let $\overline{w}_1, \overline{w}_2$ the indices of the largest TD sets included in $\Lambda(w_1)$ and $\Lambda(w_2)$ respectively, as in (3.17). It holds

$$\begin{aligned}
&\left\|\textstyle\sum_{\mathbf{p}\in\Lambda(w_2)} u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y}) - \sum_{\mathbf{p}\in\Lambda(w_1)} u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y})\right\|_{L^\infty(\Gamma;V)} = \\
&\left\|\textstyle\sum_{\mathbf{p}\in\Lambda(w_2)\setminus\Lambda(w_1)} u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y})\right\|_{L^\infty(\Gamma;V)} \leq \sum_{\mathbf{p}\in\Lambda(w_2)\setminus\Lambda(w_1)} \|u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y})\|_{L^\infty(\Gamma;V)} \leq \\
&\sum_{\mathbf{p}\notin TD(\overline{w}_1)} \|u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y})\|_{L^\infty(\Gamma;V)} = \sum_{\mathbf{p}\notin TD(\overline{w}_1)} \|u_{\mathbf{p}}(\mathbf{x})\|_V \|\mathcal{L}_{\mathbf{p}}(\mathbf{y})\|_{L^\infty(\Gamma)}\,.
\end{aligned}$$

Now use estimate (3.13) in Proposition 3.1 to bound $\|u_{\mathbf{p}}(\mathbf{x})\|_V$. Furthermore note that the $L^\infty(\Gamma)$-norm of the orthonormal Legendre polynomials can be bounded as

$$\|\mathcal{L}_{\mathbf{p}}(\mathbf{y})\|_{L^\infty(\Gamma)} = \prod_{n=1}^{N} \sqrt{2p_n + 1} \leq \left(\sqrt{3}\right)^{|\mathbf{p}|} \quad \forall \mathbf{p} \in \mathbb{N}^N,$$

so that

$$\sum_{\mathbf{p}\notin TD(\overline{w}_1)} \|u_{\mathbf{p}}(\mathbf{x})\|_V \, \|\mathcal{L}_{\mathbf{p}}(\mathbf{y})\|_{L^\infty(\Gamma)} \leq C_0 \sum_{\mathbf{p}\notin TD(\overline{w}_1)} \left(\frac{1}{\log 2}\mathbf{r}\right)^{\mathbf{p}} \frac{|\mathbf{p}|!}{\mathbf{p}!}$$

$$= C_0 \sum_{|\mathbf{p}|\geq \overline{w}_1} \left(\frac{1}{\log 2}\mathbf{r}\right)^{\mathbf{p}} \frac{|\mathbf{p}|!}{\mathbf{p}!} = C_0 \sum_{s=\overline{w}_1}^{\infty} \left(\sum_{n=1}^{N} \frac{1}{\log 2} r_n\right)^s ,$$

that tends to 0 if condition (3.18) holds, where we have exploited the generalized Newton binomial formula as in Corollary 3.1. □

Remark 3.2. *Condition* (3.18) *in Proposition 3.2 can be weakened by improving bound* (3.14)*. We recall the definition of* $I(\mathbf{p}) = \prod_{n=1}^{N} \sqrt{2p_n+1} \int_{\Gamma_n} (1-y_n^2)^{p_n} \rho(y_n) dy_n$*. Integrating* p *times by parts, one obtains*

$$\int_{-1}^{1} (1-t^2)^p \frac{1}{2} dt = \frac{2^{2p}(p!)^2}{(2p+1)!} .$$

Using Stirling's approximation formula

$$p! = \sqrt{2\pi p}\left(\frac{p}{e}\right)^p e^{\lambda_p}, \qquad \frac{1}{12p+1} \leq \lambda_p \leq \frac{1}{12p} ,$$

one can then bound

$$I(p) \leq \sqrt{\frac{\pi}{2}} \quad \Rightarrow \quad I(\mathbf{p}) \leq \left(\frac{\pi}{2}\right)^{N/2} .$$

Note that this bound is sharp, even for small values of $|\mathbf{p}|$*. Using this result rather than* (3.14) *in* (3.15) *we obtain*

$$\|u_{\mathbf{p}}\|_{V(D)} \leq C_0 \left(\frac{\pi}{2}\right)^{N/2} \left(\frac{1}{2\log 2}\mathbf{r}\right)^{\mathbf{p}} \frac{|\mathbf{p}|!}{\mathbf{p}!} \tag{3.19}$$

and, as a consequence, condition (3.18) *becomes*

$$\sum_{i=1}^{N} r_n < \frac{2\log 2}{\sqrt{3}} . \tag{3.20}$$

Note however that this is only a little improvement, being $\log 2 = 0.69$ *and* $2\log 2/\sqrt{3} = 0.80$*; moreover, since* $\pi/2 > 1$*, bound* (3.19) *does not imply that the Legendre coefficients of* u *decay regardless of the number of random variables, which was the case for the initial estimate* (3.13)*; therefore, condition* (3.20) *holds for fixed* N*, while* (3.18) *is independent of* N*.*

Following again the abstract procedure in Section 3.3.a, we substitute the estimate (3.13) in the general quasi optimal set expression (3.10). This results in the following expression for the quasi optimal polynomial sets for a general non factorizing u,

$$\Lambda(w) = \left\{ \mathbf{p} \in \mathbb{N}^N : \sum_{n=1}^{N} g_n p_n - \log \frac{|\mathbf{p}|!}{\mathbf{p}!} \leq w \right\} . \tag{3.21}$$

We refer to these sets as TD-FC sets ("TD with factorial correction" sets). We can indeed interpret the factor $\log \frac{|\mathbf{p}|!}{\mathbf{p}!}$ appearing in (3.21) as a correction factor to the TD space to take into account the intrinsic coupling between directions in the stochastic space; observe that this correction is always isotropic.

As pointed out in Remark 3.1, the quantities g_n appearing in (3.21) are better estimated numerically by a sequence of monovariate analyses: one could indeed increase the polynomial degree in one random variable at a time while keeping degree zero in all the others variables and estimate numerically the exponential rate of convergence. Observe that in such monovariate analyses the factorial term does not appear so the expected convergence rate is precisely $\sim e^{-g_p p_p}$. In the numerical results presented in the next section we have used this strategy, which seems to work particularly well.

Remark 3.3. *Observe that $\Lambda(w)$ actually depends on the number of input variables N. One can extend the definition of $\Lambda(w)$ also to the case where $\mathbf{p}$ is a sequence of natural numbers ("infinite dimensional probability space") with only a finite number of non zero terms, provided the sequence $g_n \to +\infty$ as $n \to \infty$. This is an alternative way to work with random fields, without truncating them a priori to a certain level (see e.g. [72, 21, 20]). See also [44] for a more recent adaptive algorithm in infinite dimensions.*

3.3.b Numerical Tests

In this section we show the performance of the TD-FC sets (3.21) compared to the isotropic and anisotropic versions of TD sets defined in Tables 3.1 and 3.2, as well as the best M-term approximation. We consider the following elliptic problem in one physical dimension

$$\begin{cases} -(a(x,\mathbf{y})u(x,\mathbf{y})')' = 1 & x \in D = (0,1), \mathbf{y} \in \Gamma \\ u(0,\mathbf{y}) = u(1,\mathbf{y}) = 0, & \mathbf{y} \in \Gamma \end{cases} \tag{3.22}$$

with different choices of diffusion coefficient $a(x,\mathbf{y})$, for which Assumptions 3.1 - 3.3 hold. We focus on a linear functional $\psi : V \to \mathbb{R}$ of the solution, so that $\psi(u)$ is a scalar random variable, function of $\mathbf{y}$ only. In our examples, ψ is defined as $\psi(v) = v(\frac{1}{2})$.

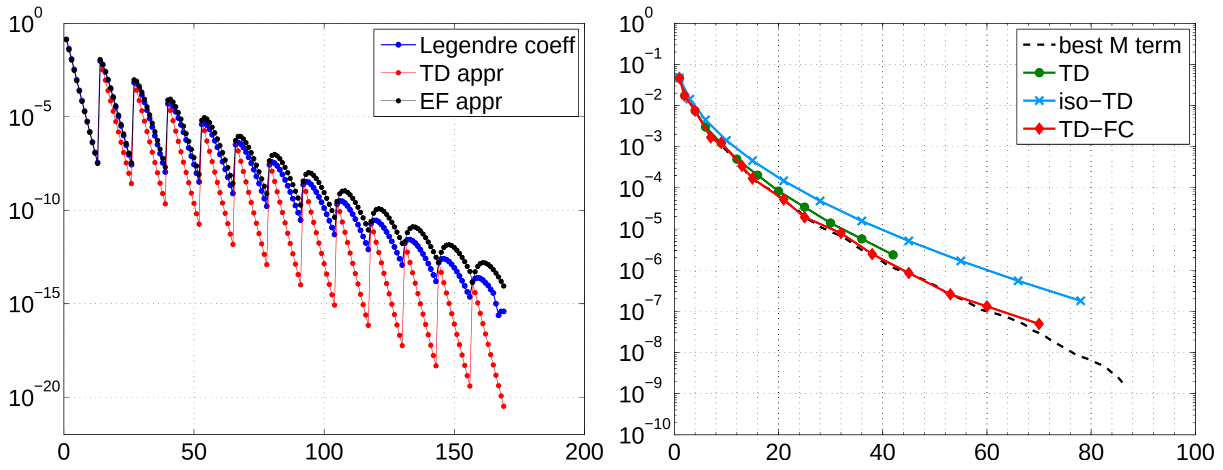
To obtain the best M-terms approximation we compute explicitly all the Legendre coefficients of $\psi(u)$ in a sufficiently large index set $\mathbb{U}$ evaluating the integrals $\psi_{\mathbf{p}} = \int_\Gamma \psi(u) L_{\mathbf{p}}(\mathbf{y})\rho(\mathbf{y})d\mathbf{y}$ with a high-level sparse grid as reference values. We order then the coefficients in decreasing order, according to their modulus, and take the first M terms of the reordered sequence as the best M-terms approximation.

The rates $\mathbf{g}$ used to build the TD-FC space, as well as the anisotropic TD space (with $\alpha_n = g_n$), are computed numerically, with a sequence of 1D analyses. For each random variable $1 \leq n \leq N$, we consider the subset $\mathbb{U}_n = \{\mathbf{p} \in \mathbb{U} : \mathbf{p}_i = 0 \text{ if } i \neq n, \mathbf{p}_n = 0, 1, 2, \ldots\}$; according to (3.13), the decay of the Legendre coefficients for this particular choice of multi-indices is $|\psi_{\mathbf{p}}| \sim e^{-g_n p_n}$, and we can then estimate the rate g_n via a linear interpolation on the quantities $\log|\psi_{\mathbf{p}}|, \mathbf{p} \in \mathbb{U}_n$.

Test 1: space independent diffusion coefficient

The first case we consider has two random variables (y_1, y_2) and a diffusion coefficient $a(x,\mathbf{y}) = 1 + 0.1y_1 + 0.5y_2$; results are shown in Figures 3.1-3.2.

Figure 3.1(a) shows the Legendre coefficients ordered in lexicographic order, giving this peculiar sawtooth shape. The first tooth corresponds to multi-indices of the form [0, k], the second one to [1, k] and so on. We have also added to the plot the estimate (3.13) in Proposition 3.1 of the magnitude of the Legendre coefficients, which leads to the TD-FC sets (3.21), as well as the estimate (3.12) which leads to the anisotropic TD spaces as in Table 3.2, with $\boldsymbol{\alpha}_n = g_n$. The plot suggests that estimate (3.13) is quite sharp, whereas the estimate corresponding to the TD space underestimates considerably the Legendre coefficients. This result highlights the importance of the



(a) Legendre coefficients in lexicographic order and their corresponding estimates based on either TD-FC or TD approximations.

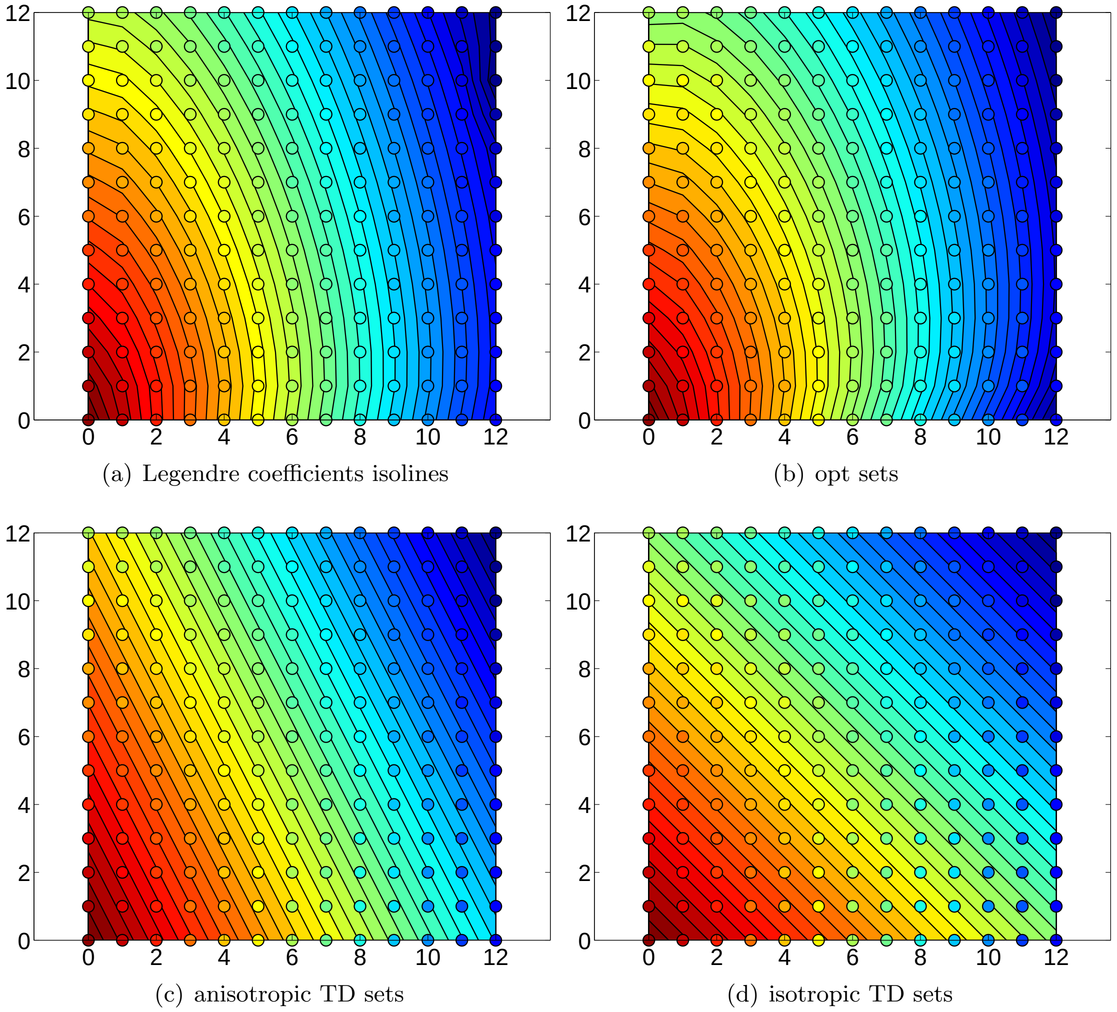
(b) Convergence of different polynomial approximations, measured as $\|\psi(u) - \psi(u_w)\|_{L^2_\rho(\Gamma)}$ versus dimension of polynomial space.

Figure 3.1: Results for $a(x, \mathbf{y}) = 1 + 0.1y_1 + 0.5y_2$. Here we have $g \simeq (2.49,\ 1.27)$, $\mathbb{U}$ =TP(12), Legendre coefficients computed with a standard Smolyak sparse grid of level 9, with Gauss-Legendre abscissae.

factorial term in (3.13). We expect, therefore, that the TD-FC approximation performs better than the aniso-TD one. Moreover, we point out the non intuitive fact that the Legendre coefficients $\psi_{\mathbf{p}}$ *are not strictly decreasing in absolute value* when listed in the lexicographic order. As an example, $|\psi_{[5\,0]}| < |\psi_{[5\,1]}|$, and the same holds for all teeth but the first few.

Figure 3.1(b) shows convergence plots for the error in L^2_ρ-norm for the various polynomial spaces used versus the dimension of the polynomial space. As the TD-FC sequence is the only sequence that captures correctly the non decreasing behaviour of the Legendre coefficients in lexicographic order, the convergence of the TD-FC sequence in Figure 3.1(b) is the closest to the best M-terms approximation, even though the anisotropic TD space give good results as well. We also point out the poor performance of the standard isotropic TD space compared to both the anisotropic TD and the TD-FC spaces: this confirms the importance of using anisotropic spaces to reduce computational costs.

It is also useful to visualize the isolines of the Legendre coefficients of the expansion of $\psi(u)$ and to compare them with the isolines corresponding to estimates (3.13) for TD-FC sets, and (3.12) for iso and aniso TD sets, see Figure 3.2. The closer the matching of the sequence of sets with the true decay of the Legendre coefficients, the faster the L^2 convergence of the approximation for ψ will be. The key property of the decay of the Legendre coefficients is the rounded shape of the isolines (see Figure 3.2(a)), properly caught only with the factorial term $\frac{|\mathbf{i}|!}{\mathbf{i}!}$ in the TD-FC set formula (Figure 3.2(b)). Also from these plots one can see the fact that the Legendre coefficients are not strictly decreasing in lexicographic order: actually close to the borders the isolines tend to bend "backward", so that for example the index [7, 1] belongs to a lower isoline than [7, 0]. However, as appears from results in Figure 3.1, approximating the isolines with "mean" straight lines as it is done in the anisotropic TD (Figure 3.2(c)) gives quite good results as well. On the other hand, using the wrong slopes for TD sets, like in isotropic TD sets (3.2(d)), will result in general in poor approximation properties.



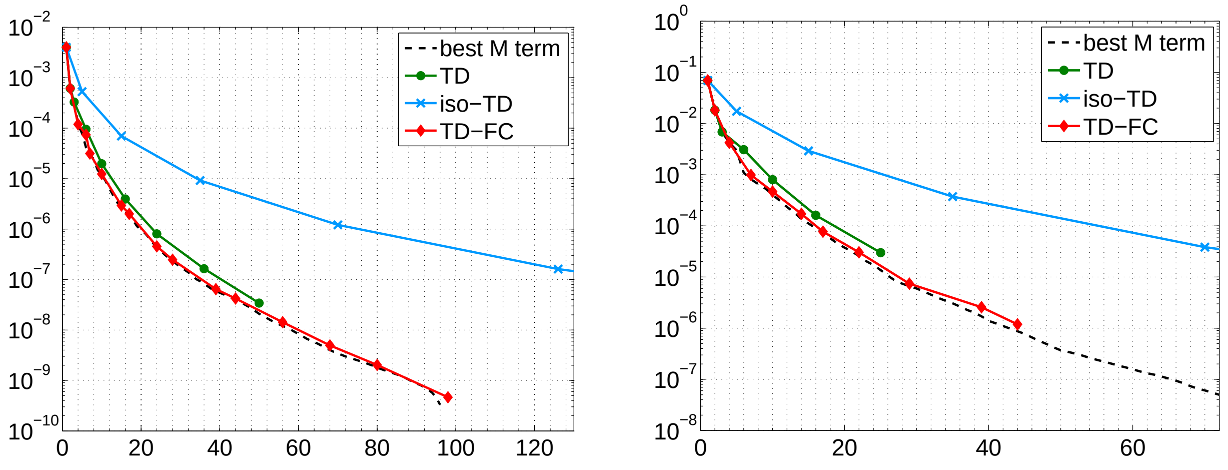
(a) Legendre coefficients isolines

(b) opt sets

(c) anisotropic TD sets

(d) isotropic TD sets

Figure 3.2: Isolines of estimated Legendre coefficients: a) true values, computed with high level sparse grids; b) estimate (3.13) leading to TD-FC sets; c) estimate (3.12) leading to aniso-TD sets with $\alpha_n = g_n$; d) estimate (3.12) with $\alpha_n = 1 \, \forall n = 1, \dots, N$, leading to standard TD sets as in Table 3.1. In all plots, each dot represents a multi-index in $\mathbb{N}^2$, and it is coloured according to the size of the corresponding exact coefficient in the Legendre expansion for ψ; on the background the isolines.



(a) Results for linear expansion. $a(x,\mathbf{y}) = 4 + y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$. Here we have $g \simeq (2.03, 4.11, 5.73, 7.05)$, reference set: $\mathbb{U} =$TD(9).

(b) Results for exponential expansion. $\log a(x,\mathbf{y}) = y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$. Here we have $g \simeq (1.95, 3.95, 5.09, 6.51)$, reference set: $\mathbb{U} =$TD(7).

Figure 3.3: Convergence of polynomial approximations for elliptic equation with the coefficient a depending also on x. Convergence measured as $\|\psi(u) - \psi(u_w)\|_{L^2_\rho(\Gamma)}$ versus the dimension of polynomial space.

Test 2: space dependent diffusion coefficient

We now consider the following two expansions:

- $a(x,\mathbf{y}) = 4 + y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$,
- $\log a(x,\mathbf{y}) = y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$.

and look at the functional $\psi(v) = v(0.7)$ (the functional $\psi(v) = v(1/2)$ is not suited for analysis in this case as, by symmetry, many of the Legendre coefficients are zero). Figure 3.3 shows the results, and again we see that the TD-FC approximation is the best performing, with anisotropic TD closely following and isotropic TD far worse.

Test 3: separable diffusion coefficient

Let us now give an example on the case of a factorizable u, as in Section 3.3.a. We recall that Section 3.3.a states that if we can express the solution $u(\mathbf{x},\mathbf{y})$ as a product $u(\mathbf{x},\mathbf{y}) = f(\mathbf{x})\prod_n v_n(y_n)$ then the Legendre coefficients can be computed as a product of 1D Legendre coefficients, and thus the optimal estimate is (3.12), leading to aniso-TD sets, rather than estimate (3.13) leading to what we have called TD-FC sets. To support our thesis, we now consider $a(\mathbf{y}) = (1 + 0.6y_1)(1 + 0.6y_2)$, so that the solution of (3.22) is $u(x,\mathbf{y}) = \frac{x(1-x)}{2a(\mathbf{y})}$.

The convergence plots for of $\psi(u)$ are shown in Figure 3.4 and confirm that in this case TD is the optimal choice, and is very close to the best M-terms approximation. Note that in this example the isotropic and anisotropic versions of TD coincide, since the two factors of v are the same.

3.3.c Alternative estimates for diffusion coefficients in exponential form

Let us consider again the model problem (3.22), with diffusion coefficient in exponential form $\log a(\mathbf{x},\mathbf{y}) = \sum_{n=1}^{N} c_n y_n$. The solution is $u(x,\mathbf{y}) = \frac{x(1-x)}{2}\frac{1}{\prod_{n=1}^{N} e^{c_n y_n}}$, therefore the solution is in

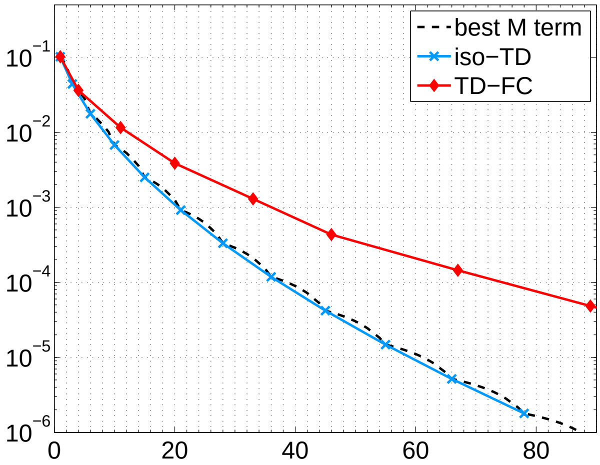


Figure 3.4: convergence of polynomial approximation of the elliptic equation (3.22) with the coefficient a that factorizes with respect to $\mathbf{y}$, $a = (1+0.6y_1)(1+0.6y_2)$. $g \simeq (1.08,\, 1.08)$, $\mathbb{U}$ =TP(12). Convergence measured as $\|\psi(u) - \psi(u_w)\|_{L^2_\rho(\Gamma)}$ versus number of Legendre coefficients (dimension of polynomial space).

separable form with $v_n = e^{c_n y_n}$; as a consequence, following the arguments in Section 3.3.a on factorizable functions, we have $\|u_{\mathbf{p}}\|_V = \|f(\mathbf{x})\|_V \prod_{n=1}^N |v_{n,p_n}|$, where v_{n,p_n} indicates the p_n-th Legendre coefficients of v_n. In this case, however, we expect the decay of v_{n,p_n} to be faster than exponential, since $v_n(y_n)$ is an entire function. Actually, the following lemma holds:

Lemma 3.1. *Given problem* (3.22) *with diffusion coefficient* $\log a(\mathbf{x},\mathbf{y}) = \sum_{n=1}^N c_n y_n$*, the* V*-norm of the Legendre coefficients of* u *can be bounded as*

$$\|u_{\mathbf{p}}\|_V \leq C_e \frac{e^{-\sum_{n=1}^N g_n p_n}}{\mathbf{p}!}, \tag{3.23}$$

with $g_n = -\log \frac{|c_n|}{\sqrt{3}}$ *and* $C_e = \|f\|_V e^{\sum_{n=1}^N |c_n|}$.

<u>Proof.</u> Since $\|u_{\mathbf{p}}\|_V = \|f(\mathbf{x})\|_V \prod_{n=1}^N |v_{n,p_n}|$ we only need to estimate $|v_{n,p_n}|$. Recalling the definition of $I(p)$ given in the proof of Proposition 3.1, one gets

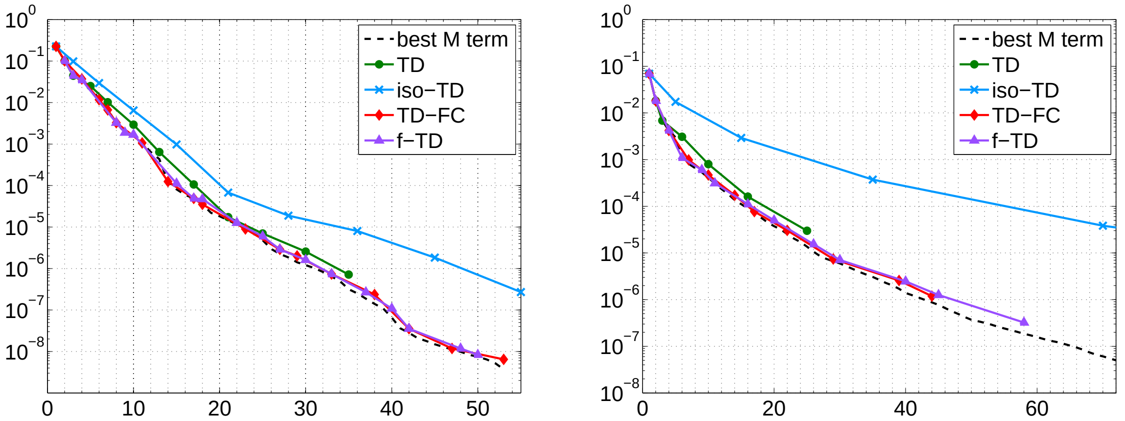
$$|v_{n,p_n}| = \left| \int_{-1}^1 L_{p_n}(y_n) v_n(y_n) \frac{dy_n}{2} \right| = \frac{\sqrt{2p_n+1}}{2^{p_n} p_n!} \left| \int_{-1}^1 e^{-c_n y_n} \left(\frac{d}{dy}\right)^{p_n} (1-y_n^2)^{p_n} \frac{dy_n}{2} \right| =$$
$$\sqrt{2p_n+1} \frac{|c_n|^{p_n}}{2^{p_n} p_n!} \int_{-1}^1 e^{-c_n y_n} (1-y_n^2)^{p_n} \frac{dy_n}{2} \leq \frac{|c_n|^{p_n} e^{|c_n|}}{2^{p_n} p_n!} I(p) \leq \frac{|c_n|^{p_n} e^{|c_n|}}{\sqrt{3}^{p_n} p_n!}.$$

The thesis follows setting $g_n = -\log \frac{|c_n|}{\sqrt{3}}$. □

Remark 3.4. *Observe that in* (3.23) *the coefficient* $u_{\mathbf{p}}$ *will tend to zero as* $|\mathbf{p}| \to \infty$ *even when* $g_n > \sqrt{3}$ *for all* $n = 1, \ldots, N$.

As a consequence, the abstract optimal space (3.10) becomes in this case

$$\Lambda(w) = \left\{ \sum_{n=1}^N p_n g_n + \sum_{n=1}^N \log(p_n!) \leq w \right\}. \tag{3.24}$$



(a) constant coefficients, $\log(a(x,\mathbf{y})+0.01) = 0.2y_1 + 2y_2$: $\tilde{g} \simeq (2.38, -0.12)$ for fTD, $g = (3.65, 1.62)$ for TD-FC and TD, reference set $\mathbb{U}$ =TP(12).

(b) sin expansion $\log a(x,\mathbf{y}) = y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$: $\tilde{g} \simeq (0.68,\ 2.81,\ 4.10,\ 5.69)$ for fTD, $g = (1.95, 3.95, 5.09, 6.51)$ for TD-FC and TD, reference set $\mathbb{U}$ =TD(7).

Figure 3.5: convergence of polynomial spaces for elliptic equation with "shifted" exponential $a(x,\mathbf{y})$, using f-TD space.

We refer to this set as anisotropic "factorial TD", or aniso-fTD in short.

We now guess that even in the more general case where $\log a(\mathbf{x},\mathbf{y}) = \sum_{n=1}^{N} c_n(\mathbf{x})y_n$ an estimate of the type of (3.23) for the Legendre coefficients of the solution holds, for some g_n, $n+1,\dots,N$. We have tested this space on two cases

- $\log(a(x,\mathbf{y})+0.01) = 0.2y_1 + 2y_2$ (constant coefficients) ;
- $\log a(x,\mathbf{y}) = y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$ (sin expansion, this one is the same as in Test 2).

Again, the rates g_n appearing in formula (3.24) can be estimated numerically with a least square approach. We will refer to these new rates as $\tilde{g_n}$ to stress the fact that they are different from the g_n we use in TD and TD-FC spaces.

The corresponding results are shown in Figure 3.5, and show that actually fTD is competing with TD-FC .

3.4 Stochastic Collocation

The Stochastic Collocation (SC) Finite Element method consists in collocating problem (3.1) in a set of points $\{\mathbf{y}_j \in \Gamma, \quad j = 1,\dots,M_w\}$, i.e. computing the corresponding solutions $u(\cdot,\mathbf{y}_j)$ and building a global polynomial approximation u_w, not necessarily interpolatory, upon those evaluations: $u_w(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{M_w} u(\mathbf{x},\mathbf{y}_j)\tilde{\psi}_j(\mathbf{y})$ for suitable multivariate polynomials $\{\tilde{\psi}_j\}_{j=1}^{M_w}$.

Building the set of evaluation points $\{\mathbf{y}_j\}$ as a cartesian product of monodimensional grids becomes quickly unfeasible, since the computational cost grows exponentially fast with the number of stochastic dimensions needed. We consider instead the so-called *sparse grid* procedure, originally introduced by Smolyak in [95] for high dimensional quadrature purposes; see also [7, 15] for polynomial interpolation. In the following we briefly review and generalize this construction.

For each direction y_n we introduce a sequence of one dimensional polynomial interpolant operators of increasing order: $\mathcal{U}_n^{m(i)} : C^0(\Gamma_n) \to \mathbb{P}_{m(i)-1}(\Gamma_n)$. Here $i \geq 1$ denotes the level of approximation and $m(i)$ the number of collocation points used to build the interpolation at level i. As a consequence, $\mathcal{U}_n^{m(i)}[q] = q$ if q is a polynomial of degree up to $m(i) - 1$. We require the function m to satisfy the following assumptions: $m(0) = 0$, $m(1) = 1$ and $m(i) < m(i + 1)$ for $i \geq 1$. In addition, let $\mathcal{U}_n^0[q] = 0$, $\forall q \in C^0(\Gamma_n)$.

Next we introduce the difference operators $\Delta_n^{m(i)} = \mathcal{U}_n^{m(i)} - \mathcal{U}_n^{m(i-1)}$, an integer value $w \geq 0$, multi-indices $\mathbf{i} \in \mathbb{N}_+^N$ and a sequence of index sets $\mathcal{I}(w)$ such that $\mathcal{I}(w) \subset \mathcal{I}(w + 1)$ and $\mathcal{I}(0) = \{(1, 1, \ldots, 1)\}$. We define the sparse grid approximation of $u : \Gamma \to V$ at level w as

$$u_w(\mathbf{y}) = \mathcal{S}_{\mathcal{I}(w)}^m[u](\mathbf{y}) = \sum_{\mathbf{i} \in \mathcal{I}(w)} \bigotimes_{n=1}^{N} \Delta_n^{m(i_n)}[u](\mathbf{y}). \tag{3.25}$$

As pointed out in [41], it is desirable that the sum (3.25) has some telescopic properties. To ensure this we have to impose some additional constraints on $\mathcal{I}$. Following [41] we say that a set $\mathcal{I}$ is *admissible* if $\forall\, \mathbf{i} \in \mathcal{I}$

$$\mathbf{i} - \mathbf{e}_j \in \mathcal{I} \text{ for } 1 \leq j \leq N, i_j > 1. \tag{3.26}$$

We refer to this property as *admissibility condition*, or *ADM* in short. Given a set $\mathcal{I}$ we will denote by $\mathcal{I}^{ADM}$ the smallest set such that $\mathcal{I} \subset \mathcal{I}^{ADM}$ and $\mathcal{I}^{ADM}$ is admissible.

It is now possible to rewrite (3.25) in terms of linear combinations of tensor grids interpolations:

$$u_w(\mathbf{y}) = \sum_{\mathbf{i} \in \mathcal{I}(w)^{ADM}} c_{\mathbf{i}} \bigotimes_{n=1}^{N} \mathcal{U}_n^{m(i_n)}[u](\mathbf{y}), \quad c_{\mathbf{i}} = \sum_{\substack{\mathbf{j} = \{0,1\}^N: \\ \mathbf{i}+\mathbf{j} \in \mathcal{I}(w)^{ADM}}} (-1)^{|\mathbf{j}|}. \tag{3.27}$$

Observe that many coefficients $c_{\mathbf{i}}$ in (3.27) are zero. The set of all evaluation points needed is called *sparse grid* and denoted by $\mathcal{H}_{\mathcal{I}(w)}^m \subset \Gamma$ (see Figure 3.6). We also introduce the tensor notation

$$m(\mathbf{i}) = \prod_{n=1}^{N} m(i_n), \quad \Delta^{m(\mathbf{i})}[u] = \bigotimes_{n=1}^{N} \Delta^{m(i_n)}[u], \quad \mathcal{U}^{m(\mathbf{i})}[u] = \bigotimes_{n=1}^{N} \mathcal{U}^{m(i_n)}[u].$$

To fully characterize the sparse approximation operator $\mathcal{S}_{\mathcal{I}(w)}^m$ introduced in (3.25) one has to provide the sequence of sets $\mathcal{I}(w)$, the relation $m(i)$ between the level i and the number of points in the corresponding one dimensional polynomial interpolation formula $\mathcal{U}^{m(i)}$, and the family of points to be used at each level, e.g. Clenshaw-Curtis or Gauss abscissae (see e.g. [102]).

In what follows we will consider Clenshaw-Curtis abscissae and the “doubling” rule $m(i) = db(i)$,

$$db(i) = \begin{cases} 0 \text{ if } i = 0 \\ 1 \text{ if } i = 1 \\ 2^{i-1} + 1, \text{ if } i > 1, \end{cases} \tag{3.28}$$

which leads to nested grids. The classical Smolyak sparse grid (SM) uses $\mathcal{I}(w) = \{\mathbf{i} \in \mathbb{N}_+^N : |\mathbf{i} - \mathbf{1}| \leq w\}$, which clearly satisfies the admissibility condition (3.26). A quasi optimal choice of $\mathcal{I}(w)$ will be discussed in the next Section.

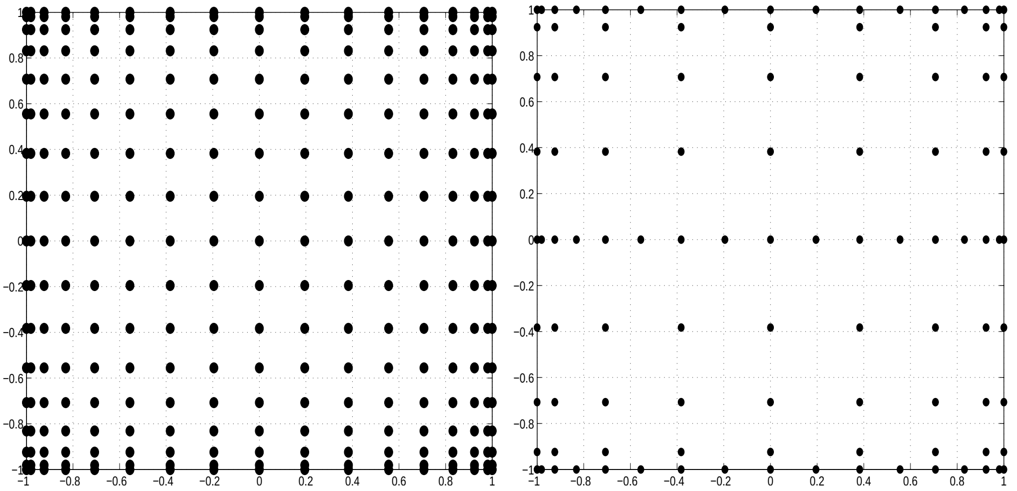


Figure 3.6: comparison between a tensor grid (left) and the TD-FC sparse grid (right) derived with the procedure explained in Section 3.4.a.

3.4.a Quasi-optimal sparse grids

We now aim at constructing the quasi-optimal sparse grid for the Stochastic collocation method, i.e. we aim at choosing the best sequence of sets of indices. Let us define the error associated to a sparse grid as

$$E(\mathcal{S}^m_{\mathcal{I}(w)}) = \left\| u - \mathcal{S}^m_{\mathcal{I}(w)}[u] \right\|_{V\otimes L^2_\rho(\Gamma)},$$

and the work $W(\mathcal{S})$ as the number of evaluations needed, i.e.

$$W(\mathcal{S}^m_{\mathcal{I}(w)}) = |\mathcal{H}^m_{\mathcal{I}(w)}|.$$

Our goal is then to find the optimal set $\mathcal{S}$ that minimizes the error with a total work smaller or equal to a maximum work W, or alternatively the set that minimizes the work with an error smaller than or equal to a given threshold ϵ. This is a classical knapsack problem and we adopt a greedy algorithm to solve it. To this end we define the error and work contribution of a multi-index $\mathbf{i}$. Let $\mathcal{J}$ be any set of indices such that $\mathbf{i} \notin \mathcal{J}$ and $\{\mathcal{J} \cup \mathbf{i}\}$ is admissible. Then the error contribution of $\mathbf{i}$ is

$$\Delta E(\mathbf{i}) = \left\| \mathcal{S}^m_{\{\mathcal{J}\cup\mathbf{i}\}}[u] - \mathcal{S}^m_{\mathcal{J}}[u] \right\|_{V\otimes L^2_\rho(\Gamma)} \tag{3.29}$$

and the work contribution is

$$\Delta W(\mathbf{i}) = |W(\mathcal{S}^m_{\{\mathcal{J}\cup\mathbf{i}\}}) - W(\mathcal{S}^m_{\mathcal{J}})|. \tag{3.30}$$

Observe that the error contribution defined in (3.29) is always independent of the set $\mathcal{J}$, since indeed

$$\Delta E(\mathbf{i}) = \left\| \sum_{\mathbf{j}\in\{\mathcal{J}\cup\mathbf{i}\}} \Delta^{m(\mathbf{j})}[u] - \sum_{\mathbf{j}\in\{\mathcal{J}\}} \Delta^{m(\mathbf{j})}[u] \right\|_{V\otimes L^2_\rho(\Gamma)} = \left\| \Delta^{m(\mathbf{i})}[u] \right\|_{V\otimes L^2_\rho(\Gamma)}. \tag{3.31}$$

On the other hand, the work contribution (3.30) will depend in general on the set $\mathcal{J}$, except in the case of nested abscissae, as for Clenshaw Curtis nodes, which is the case considered here. In this case indeed the evaluation of the extra term $\Delta^{m(\mathbf{i})}[u] = \bigotimes_{n=1}^N (\mathcal{U}^{m(i_n)} - \mathcal{U}^{m(i_n-1)})[u]$ implies evaluations only in the extra points added at level i_n in each direction, irrespectively of the set $\mathcal{J}$, provided that $\mathcal{J}$ is admissible.

Following [15, 41] we can now define the profit of an index $\mathbf{i}$ as

$$P(\mathbf{i}) = \frac{\Delta E(\mathbf{i})}{\Delta W(\mathbf{i})}$$

and identify the optimal sparse approximation operator $\mathcal{S}^*$ as the one using the set of most profitable indices, i.e. $\mathcal{I}^*(\epsilon) = \{\mathbf{i} \in \mathbb{N}_+^N : P(\mathbf{i}) \geq \epsilon\}$.

To build the set $\mathcal{I}^*$ we rely on sharp estimates for both $\Delta E(\mathbf{i})$ and $\Delta W(\mathbf{i})$. Since, using Clenshaw-Curtis abscissae and the doubling rule $db(\cdot)$, we get nested grids, we can compute *exactly* $\Delta W(\mathbf{i})$ as

$$\Delta W(\mathbf{i}) = \prod_{n=1}^{N} (db(i_n) - db(i_n - 1)), \tag{3.32}$$

with $db(i_n)$ as in (3.28).

On the other hand, deriving a rigorous bound for $\Delta E(\mathbf{i})$ is not as easy. For instance, through numerical investigations on the model function $f(y_1, y_2) = \frac{1}{1+c_1y_1+c_2y_2}$, one can conjecture the size of a generic $\Delta E(\mathbf{i})$ to be closely related to the norm of the corresponding Legendre coefficient $f_{m(\mathbf{i-1})}$, with a correcting factor due to the interpolation operator norm. To be more precise, we conjecture the following estimate for $\Delta E(\mathbf{i})$,whenever f is an analytic function:

$$\Delta E(\mathbf{i})[f] \lesssim \left\| f_{m(\mathbf{i}-1)} \right\|_V \textstyle\prod_{n=1}^{N} \mathbb{L}_n^{m(i_n)}, \tag{3.33}$$

where $a \lesssim b$ means that there exists a constant c independent of $\mathbf{i}$ such that $a \leq cb$ and $\mathbb{L}_n^{m(i)}$ is the Lebesgue constant for the interpolation operator $\mathcal{U}_n^{m(i)}$, defined as

$$\mathbb{L}_n^{m(i)} = \sup_{v \in C^0(\Gamma_n)} \frac{\left\| \mathcal{U}_n^{m(i)} v \right\|_{L^\infty(\Gamma_n)}}{\|v\|_{L^\infty(\Gamma_n)}}. \tag{3.34}$$

For Clenshaw-Curtis abscissae with doubling relation the Lebesgue constant can be shown to be

$$\mathbb{L}(db(i)) \leq \frac{2}{\pi} \log(db(i_n) + 1) + 1,$$

see e.g. [28, 29]. Figure 3.7 shows the quality of estimate (3.33), and numerical results in the next Section also confirm that such an estimate is accurate enough for our purposes.

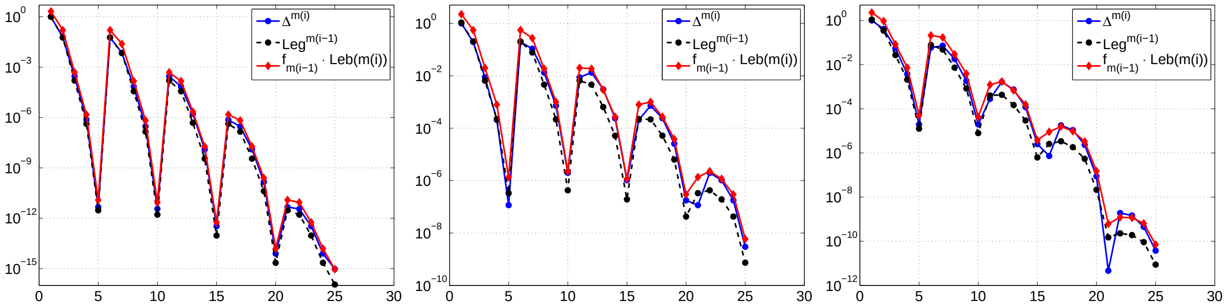
Starting from (3.32) and (3.33), we can estimate the profit of each index, and estimate the sequence $\mathcal{S}_{\mathcal{I}^*(\epsilon)}$ of quasi-optimal grids with

$$\mathcal{I}^*(\epsilon) = \left\{ \mathbf{i} \in \mathbb{N}_+^N : \frac{C_0 \exp\left(-\sum_{n=1}^{N} db(i_n - 1) g_n \right) \frac{|db(\mathbf{i}-\mathbf{1})|!}{db(\mathbf{i}-\mathbf{1})!} \prod_{n=1}^{N} \mathbb{L}_n^{m(i_n)}}{\prod_{n=1}^{N} (db(i_n) - db(i_n - 1))} \geq \epsilon \right\}^{ADM} \tag{3.35}$$

with $\epsilon > 0 \in \mathbb{R}$. Equivalently, for $w = 0, 1, \ldots$ we can define the sequence of sets

$$\mathcal{I}^*(w) = \left\{ \mathbf{i} \in \mathbb{N}_+^N : \sum_{i=n}^{N} db(i_n - 1) g_n - \log \frac{|db(\mathbf{i}-\mathbf{1})|!}{db(\mathbf{i}-\mathbf{1})!} - \sum_{n=1}^{N} \log \frac{\frac{2}{\pi} \log(db(i_n) + 1) + 1}{db(i_n) - db(i_n - 1)} \leq w \right\}^{ADM} \tag{3.36}$$

that will be used in (3.25) to build the quasi optimal sparse grids. We will refer to these "quasi best M-terms grids" as EW grids ("Error-Work" grids).



(a) $f = \frac{1}{1+0.1y_1+0.1y_2}$ (b) $f = \frac{1}{1+0.3y_1+0.3y_2}$ (c) $f = \frac{1}{1+0.1y_1+0.5y_2}$

Figure 3.7: Numerical comparison between $\Delta E(\mathbf{i})$ and $|u_{m(\mathbf{i-1})}|$ for a scalar function u of the form $f(y_1, y_2) = \frac{1}{1+c_1y_1+c_2y_2}$. Both the $\Delta E(\mathbf{i})$ for $\mathbf{i} \in TP(4)$ and the corresponding Legendre coefficients $|f_{m(\mathbf{i}-1)}|$ have been computed with a standard sparse grid SM(10).

Remark 3.5. *Observe that estimate* (3.32) *of the work* $\Delta W(\mathbf{i})$ *associated to a multi-index* $\mathbf{i}$ *is valid only if the underlying set of multi-indices is admissible. This is why in formulae* (3.35) *and* (3.36) *we have explicitly enforced the admissibility condition in the construction of the optimal set. In practice, this simply implies that if at level* w *an index* $\mathbf{j}$ *is added, all indices* $\left\{\mathbf{i} \in \mathbb{N}^N : i_1 \leq j_1, i_2 \leq j_2, \ldots, i_N \leq j_N\right\}$ *have to be added as well, if not already present in the set. Note that such operation is not much demanding from a computational point of view.*

3.4.b Numerical tests on sparse grids

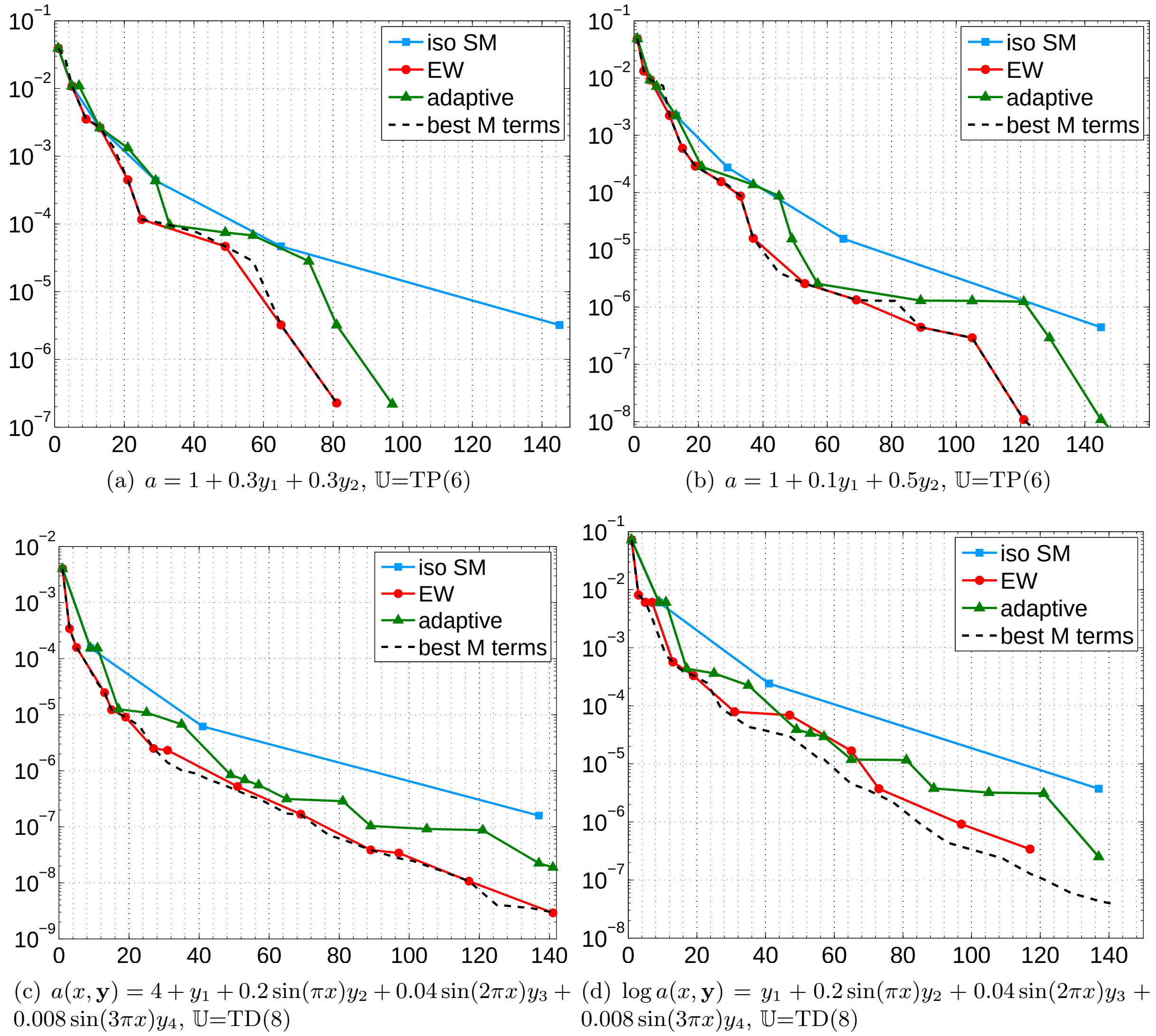
In this Section we consider the same problem as in Section 3.3.b and use it to test the performance of the EW grids derived above, comparing them with the classical SM grid and the best M-terms approximation.

To approximate the best M-terms we again consider a sufficiently large set $\mathbb{U}$ of multi-indices and for each of them we compute $\Delta W(\mathbf{i})$, $\Delta E(\mathbf{i})$ and their profit $P(\mathbf{i})$. Next, we sort the multi-indices according to $P(\mathbf{i})$, modify the sequence to fulfil the ADM condition (3.26) and compute the sparse grids according to this sequence.

We remark that the procedure just described only leads to an approximation of the best M-terms solution. Indeed, on the one hand replacing the total error $E(\mathcal{S})$ with the sum $\sum_{\mathbf{i}} \Delta E(\mathbf{i})$ provides only an upper bound that could be pessimistic because of possible cancellations, since the details $\Delta^{m(\mathbf{i})}[u]$ are not mutually orthogonal, in general. On the other hand, the fact that the most profitable index may be not admissible suggests that the solution cannot be found with a greedy algorithm. Here the coefficients g_n in (3.36) are estimated numerically as in Section 3.3.b.

We also compare our results with the dimension adaptive algorithm [41], in the implementation proposed in [56] and available at `http://www.ians.uni-stuttgart.de/spinterp`. This is an adaptive algorithm that given a sparse grid $\mathcal{S}_{\mathcal{I}}$ explores all neighbour multi-indices and adds to $\mathcal{I}$ the most profitable one. The algorithm implemented in [56] has a tunable parameter $\widetilde{\omega}$ that allows one to move continuously from the classical Smolyak formula ($\widetilde{\omega} = 0$) to the fully adaptive algorithm ($\widetilde{\omega} = 1$). Following [56], in the present work we have set $\widetilde{\omega} = 0.9$, that numerically has been proved to be a good performing choice. The cost of this algorithm is the *total* number of evaluations needed, including also those necessary to explore all neighbours, to find the most profitable multi-index.

Figure 3.8 shows the convergence of the quantity $\|\psi(u) - \psi(u_w)\|_{L^2_\rho(\Gamma)}$ versus the number of



(a) $a = 1 + 0.3y_1 + 0.3y_2$, $\mathbb{U}$=TP(6)

(b) $a = 1 + 0.1y_1 + 0.5y_2$, $\mathbb{U}$=TP(6)

(c) $a(x, \mathbf{y}) = 4 + y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$, $\mathbb{U}$=TD(8)

(d) $\log a(x, \mathbf{y}) = y_1 + 0.2\sin(\pi x)y_2 + 0.04\sin(2\pi x)y_3 + 0.008\sin(3\pi x)y_4$, $\mathbb{U}$=TD(8)

Figure 3.8: Results for EW sparse grids compared with best M-terms , isotropic Smolyak and dimension adaptive algorithm. Convergence is measured as $\|\psi(u) - \psi(u_w)\|_{L^2(\Gamma)}$ versus number of evaluations (grid points).

grid points, for the different sparse grids considered. The L^2_ρ-norm has been computed with a high level isotropic Smolyak grid. The EW grid is the best performing, even compared to the a-posteriori dimension adaptive algorithm implemented in [56], and the closest to the best M-terms grids sequence.

Remark 3.6. *A similar approach, based on estimates for ΔE and ΔW is possible also for the case of not nested grid points, as for the Gauss-Legendre quadrature points. However, in this case the estimate of ΔW is "path dependent" and any "path independent" estimate will be too pessimistic to build effective index sets.*

3.5 Conclusions

In this chapter we have proposed a new sequence of polynomial subspaces (TD-FC spaces in short) to be used in the solution of elliptic stochastic PDEs with Stochastic Galerkin method in the case of a solution that depends analytically on all random variables. The new polynomial spaces are based on sharp estimates of the decay of the Legendre coefficients.

The performances of TD-FC spaces have been assessed through some simple test cases. Here we have compared TD-FC with some standard choices of polynomial spaces and with the best M-terms approximation of the solution, that can be explicitly built for the examples considered. Results show that the TD-FC spaces perform better than the standard anisotropic TD ones, and are close to the best M-terms approximation a clear indication that our estimates of the decay of the Legendre coefficients are sharp. However, standard spaces may still have reasonable performances, if used in an appropriate anisotropic framework.

Using the estimate for the decay of the Legendre coefficients we have also defined a new class of sparse grids to be used in the context of Stochastic Collocation, relying on the concept of profit of each multi-index in the sparse grid. Again numerical tests show that these new sparse grids outperform the classical Smolyak construction and perform better than the a-posteriori dimension adaptive algorithm proposed in [41] (see also [56]). The reason for this appearent success is that our algorithm picks up the hierarchical surpluses based purely on a priori estimates and inexpensive y-one dimensional auxiliary problems. These estimates turn out to be quite sharp, and do not have any extra cost to explore neighbor points as the algorithm in [56] does.

The new polynomial spaces and sparse grids proposed here are valid in the case of analytic dependence of the solution on the random variables. We point out, however, that the general strategy outlined in Sections 3.3.a and 3.4.a on how to build optimal polynomial spaces / sparse grids, is applicable to any problem and any kind of underlying random variables. Of course, this strategy requires a sharp estimate of the decay of the coefficients of the spectral expansion of the solution on a orthonormal hierarchical basis (not necessarily polynomial). This step is highly problem dependent and should be analyzed carefully in each situation, as we did here for a linear elliptic PDE with a stochastic coefficient dependent on uniformly distributed random variables.

Appendix

Proof of Theorem 3.1

Let us consider two sufficiently smooth N-dimensional functions $f(\mathbf{y}), g(\mathbf{y}) : \mathbb{R}^N \to \mathbb{R}$; an index $i \in \mathbb{N}$, $1 \leq i \leq N$; a set $\mathcal{S}$ of indices with cardinality $\mathscr{S}$; a multi-index $\mathbf{s} \in \mathbb{N}^N$. We use the following notation:

- $\partial_i f$ denotes the derivative of f in the i-th direction: $\partial_i f = \frac{\partial}{\partial y_i} f$;

- $\partial_{\mathcal{S}} f$ denotes the $\mathscr{S}$-th order mixed derivative of f with respect to all the directions included in $\mathcal{S}$. As an example, if $\mathcal{S} = \{1\,1\,2\,4\,4\,4\}$ then

$$\partial_{\mathcal{S}} f = \partial_{1\,1\,2\,4\,4\,4} f = \frac{\partial^6}{\partial_{y_1}\partial_{y_1}\partial_{y_2}\partial_{y_4}\partial_{y_4}\partial_{y_4}} f = \frac{\partial^6}{\partial^2_{y_1}\partial_{y_2}\partial^3_{y_4}} f.$$

- $\mathbf{s}$ is the multi-index corresponding to the set $\mathcal{S}$ such that $\partial_{\mathcal{S}} f = \partial_{\mathbf{s}} f$. In the previous example $\mathbf{s} = [2\,1\,0\,3]$ is the multi-index corresponding to the set $\mathcal{S} = \{1\,1\,2\,4\,4\,4\}$.

Lemma 3.2 (generalized Leibniz rule)**.** *Given a set of indices $\mathcal{K}$ with cardinality $\mathscr{K}$ and two functions $f, g : \mathbb{R}^N \to \mathbb{R}$, $f, g \in \mathcal{C}^{\mathscr{K}}(\mathbb{R}^N)$,*

$$\partial_{\mathcal{K}}(fg) = \sum_{\mathcal{S}\in\mathcal{P}(\mathcal{K})} \frac{\partial^{\mathscr{S}} f}{\prod_{i\in\mathcal{S}} \partial y_i} \frac{\partial^{\mathscr{K}-\mathscr{S}} g}{\prod_{i\notin\mathcal{S}} \partial y_i} = \sum_{\mathcal{S}\in\mathcal{P}(\mathcal{K})} \partial_{\mathcal{S}} f \, \partial_{\mathcal{K}\setminus\mathcal{S}} g, \tag{3.37}$$

where $\mathcal{P}(\mathcal{K})$ represents the power set of $\mathcal{K}$.

Lemma 3.3. *Let $a(\mathbf{x}, \mathbf{y})$ be a diffusion coefficient for equation* (3.1) *that satisfies Assumptions 3.1 - 3.3. Then the derivatives of u can be bounded as*

$$\|\partial_{\mathbf{k}} u(\mathbf{y})\|_V \leq C_0 d_{|\mathbf{k}|} \mathbf{r}^{\mathbf{k}} \quad \forall \mathbf{y} \in \Gamma,$$

where $C_0 = \frac{\|f\|_{V'}}{a_{min}}$, $\mathbf{r}$ as in Assumption 3.3, and $\{d_n\}_{n\in\mathbb{N}}$ is a sequence defined as:

$$d_0 = 1,\ d_n = \sum_{i=0}^{n-1} \binom{n}{i} d_i. \tag{3.38}$$

Proof.

We start by rewriting the statement using the correspondence between $\mathbf{k}$ and its equivalent set $\mathcal{K}$

$$\|\partial_{\mathbf{k}} u(\cdot, \mathbf{y})\|_V = \|\partial_{\mathcal{K}} \nabla u(\cdot, \mathbf{y})\|_{L^2(D)} \leq C_0 d_{\mathscr{K}} \mathbf{r}^{\mathbf{k}}, \quad \forall \mathbf{y} \in \Gamma.$$

We will first prove something closely related, namely

$$\left\| \sqrt{a(\cdot, \mathbf{y})} \partial_{\mathcal{K}} \nabla u(\cdot, \mathbf{y}) \right\|_{L^2(D)} \leq \frac{\|f\|_{V'}}{\sqrt{a_{min}}} d_{\mathscr{K}} \mathbf{r}^{\mathbf{k}} \quad \forall \mathbf{y} \in \Gamma, \tag{3.39}$$

from which the previous inequality follows immediately. Let us start with a weak formulation of (3.1) in the physical space only, i.e.

Find $u \in V \otimes L^2_\rho(\Gamma)$ such that for almost every $\mathbf{y} \in \Gamma$ it holds

$$\int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x} = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x} \quad \forall v \in V. \tag{3.40}$$

According to Lemma 3.2, the $\partial_{\mathcal{K}}$ derivative of this weak formulation with respect to $\mathbf{y}$ is

$$\int_D \sum_{\mathcal{S}\in\mathcal{P}(\mathcal{K})} \partial_{\mathcal{S}} \nabla u(\mathbf{x}, \mathbf{y}) \partial_{\mathcal{K}\setminus\mathcal{S}} a(\mathbf{x}, \mathbf{y}) \nabla v(\mathbf{x}) d\mathbf{x} = 0,$$

and putting in evidence the $\partial_{\mathcal{K}} \nabla u$ term

$$\int_D a(\mathbf{x}, \mathbf{y}) \partial_{\mathcal{K}} \nabla u(\mathbf{x}, \mathbf{y}) \nabla v(\mathbf{x}) d\mathbf{x} = -\int_D \sum_{\mathcal{S}\in\mathcal{P}(\mathcal{K}), \mathcal{S}\neq\mathcal{K}} \partial_{\mathcal{S}} \nabla u(\mathbf{x}, \mathbf{y}) \partial_{\mathcal{K}\setminus\mathcal{S}} a(\mathbf{x}, \mathbf{y}) \nabla v(\mathbf{x}) d\mathbf{x}.$$

Next we choose $v = \partial_{\mathcal{K}} u$ and use Cauchy-Schwarz inequality on the right hand side:

$$\left\| \sqrt{a(\cdot,\mathbf{y})}\, \partial_{\mathcal{K}} \nabla u(\cdot,\mathbf{y}) \right\|^2_{L^2(D)} \leq$$
$$\sum_{\mathcal{S}\in\mathcal{P}(\mathcal{K}),\, \mathcal{S}\neq\mathcal{K}} \left\| \frac{\partial_{\mathcal{K}\setminus\mathcal{S}}\, a}{a}(\cdot,\mathbf{y}) \right\|_{L^\infty(D)} \left\| \sqrt{a(\cdot,\mathbf{y})} \partial_{\mathcal{S}} \nabla u(\cdot,\mathbf{y}) \right\|_{L^2(D)} \left\| \sqrt{a(\cdot,\mathbf{y})} \partial_{\mathcal{K}} \nabla u(\cdot,\mathbf{y}) \right\|_{L^2(D)}.$$

Now simplify $\left\| \sqrt{a(\cdot,\mathbf{y})} \partial_{\mathcal{K}} \nabla u(\cdot,\mathbf{y}) \right\|_{L^2(D)}$ on both sides and reorder the sum on the right hand side according to the cardinality of the subsets $\mathcal{S}$. From here on we omit the dependence of a and u on $\mathbf{x}$, $\mathbf{y}$, to have a lighter notation. We have

$$\left\| \sqrt{a}\, \partial_{\mathcal{K}} \nabla u \right\|_{L^2(D)} \leq \sum_{i=0}^{\mathscr{K}-1} \sum_{\mathcal{S}\in\mathcal{P}(\mathcal{K}), \mathscr{S}=i} \left\| \frac{\partial_{\mathcal{K}\setminus\mathcal{S}} a}{a} \right\|_{L^\infty(D)} \left\| \sqrt{a} \partial_{\mathcal{S}} \nabla u \right\|_{L^2(D)}. \tag{3.41}$$

We are finally in position to prove (3.39). We will proceed by induction on (3.39), using (3.41) and Assumption 3.3 on the decay of a.

Case $\mathscr{K} = 0$. In this case (3.39) reads

$$\left\| \sqrt{a} \nabla u \right\|_{L^2(D)} \leq \frac{\|f\|_{V'}}{\sqrt{a_{min}}} d_0,$$

which is true setting $d_0 = 1$.

Case $\mathscr{K} = 1$. If $\mathcal{K} = \{j\}$, $1 \leq j \leq N$, (3.39) reads

$$\left\| \sqrt{a} \partial_j \nabla u \right\|_{L^2(D)} \leq \frac{\|f\|_{V'}}{\sqrt{a_{min}}} d_1 r_j = \frac{\|f\|_{V'}}{\sqrt{a_{min}}} r_j \binom{1}{0} d_0 = \frac{\|f\|_{V'}}{\sqrt{a_{min}}} r_j d_0 = \frac{\|f\|_{V'}}{\sqrt{a_{min}}} r_j.$$

To prove this, consider (3.41). Using Assumption 3.3 and the result for case $\mathscr{K} = 0$ one has precisely

$$\left\| \sqrt{a} \partial_j \nabla u \right\|_{L^2(D)} \leq \left\| \frac{\partial_j a}{a} \right\|_{L^\infty(D)} \left\| \sqrt{a} \nabla u \right\|_{L^2(D)} \leq r_j \frac{\|f\|_{V'}}{\sqrt{a_{min}}}.$$

General $\mathscr{K}$. Consider now a general $\mathcal{K}$, and suppose (3.39) holds for any set $\mathcal{S}$ with cardinality $\mathscr{K} - 1$. Use this induction hypothesis and again Assumption 3.3 on (3.41), denoting with $\mathbf{s}$ the multi-index corresponding to the set $\mathcal{S}$ and with $\mathbf{s}*$ the multi-index corresponding to the set $\mathcal{K} \setminus \mathcal{S}$. This yields

$$\left\| \sqrt{a} \partial_{\mathcal{K}} \nabla u \right\|_{L^2(D)} \leq \sum_{i=0}^{\mathscr{K}-1} \sum_{\mathcal{S}\in\mathcal{P}(\mathcal{K}), \mathscr{S}=i} \mathbf{r}^{\mathbf{s}*} \frac{\|f\|_{V'}}{\sqrt{a_{min}}} d_{\mathscr{S}} \mathbf{r}^{\mathbf{s}}.$$

Next note that:

$$\mathbf{r}^{\mathbf{s}*}\mathbf{r}^{\mathbf{s}} = \prod_{j\in\mathcal{K}\setminus\mathcal{S}} r_j \prod_{j\in\mathcal{S}} r_j = \prod_{j\in\mathcal{K}} r_j = \mathbf{r}^{\mathbf{k}},$$

and that the number of subsets $\mathcal{S}$ with cardinality i is $\binom{\mathscr{K}}{i}$. Then

$$\left\| \sqrt{a} \partial_{\mathcal{K}} \nabla u \right\|_{L^2(D)} \leq \frac{\|f\|_{V'}}{\sqrt{a_{min}}} \mathbf{r}^{\mathbf{k}} \sum_{i=0}^{\mathscr{K}-1} \binom{\mathscr{K}}{i} d_i = \frac{\|f\|_{V'}}{\sqrt{a_{min}}} \mathbf{r}^{\mathbf{k}} d_{\mathscr{K}}$$

which proves the result.

□

Lemma 3.4. *The sequence* $\{d_n\}_{n\in\mathbb{N}}$ *defined in* (3.38) *can be bounded as*

$$d_n \leq \left(\frac{1}{\log 2}\right)^n n! \tag{3.42}$$

Proof. From definition (3.38) we have

$$d_n = \sum_{i=0}^{n-1} \binom{n}{i} d_i = \sum_{i=0}^{n-1} \frac{n!}{i!(n-i)!} d_i \,.$$

Let $f_n = \dfrac{d_n}{n!}$; the recurrency relation then becomes

$$f_n = \sum_{i=0}^{n-1} \frac{f_i}{(n-i)!}\ , \quad f_0 = f_1 = 1. \tag{3.43}$$

We now show by induction that $f_n \leq C\alpha^n$, with $C, \alpha \in \mathbb{R}$. Enforcing $1 = f_0 \leq C$ and $1 = f_1 \leq C\alpha$ results in $C \geq 1$ and $\alpha \geq 1$. Next, we reorder the sum in (3.43) and exploit the inductive hypothesis:

$$f_n = \sum_{i=0}^{n-1} \frac{f_{n-1-i}}{(1+i)!} \leq \sum_{i=0}^{n-1} \frac{C\alpha^{n-1-i}}{(1+i)!} = C\alpha^n \sum_{i=0}^{n-1} \frac{\alpha^{-(1+i)}}{(1+i)!} = C\alpha^n \left(e^{\frac{1}{\alpha}} - 1\right) \leq C\alpha^n,$$

where the last inequality holds true provided we choose $e^{\frac{1}{\alpha}} - 1 \leq 1$. Therefore we take $\alpha = (\log 2)^{-1}$ and $C = 1$, yielding $f_n \leq (\log 2)^{-n}$ and $d_n \leq (\log 2)^{-n}\, n!$

□

Theorem 3.1. *Let* $a(\mathbf{x}, \mathbf{y})$ *be a diffusion coefficient for equation* (3.1) *that satisfies Assumptions 3.1 - 3.3. Then the derivatives of* u *can be bounded as*

$$\|\partial_{\mathbf{i}} u(\mathbf{y})\|_V \leq C_0 |\mathbf{i}|!\, \tilde{\mathbf{r}}^{\mathbf{i}} \quad \forall \mathbf{y} \in \Gamma.$$

Here $C_0 = \dfrac{\|f\|_{V'}}{a_{min}}$ *and* $\tilde{\mathbf{r}} = \left(\dfrac{1}{\log 2}\right)\mathbf{r}$, *with* $\mathbf{r}$ *as in Assumption 3.3.*

Proof. Combine Lemma 3.3 and 3.4. □

Chapter 4

Application of optimal sparse grids to groundwater flow problems

4.1 Introduction

The motion of fluids in porous media can be described by the well-known Darcy's equations, see e.g. [8], which prescribe a proportionality relation between fluid flux and pressure gradient that drives the flow ("pressure head " in the hydrology literature). The proportionality constant is the so-called permeability, which is a physical property of the porous medium measuring indeed the tendency of the medium to let fluids pass through it.

Denoting by D the computational domain, p the water pressure, a the permeability field, Φ the water flux and f the forcing term acting on the system, the Darcy's equations read:

$$\begin{cases} -a\nabla p = \Phi & \text{in } D, \\ \operatorname{div}(\Phi) = f & \text{in } D, \end{cases} \tag{4.1}$$

endowed with suitable boundary conditions. The first equation is the constitutive law stating the proportionality between flux and pressure, and it is the analog to Fourier's law for heat transfer or Ohm's law for electrical circuits, while the second one states the conservation of the fluid flux. In the following, we will refer to (4.1) as the "mixed" formulation of the Darcy problem. The mixed formulation can be trivially recast as an elliptic PDE

$$-\operatorname{div}(a\nabla p) = f \quad \text{in } D, \tag{4.2}$$

again to be complemented with suitable boundary conditions. This formulation will be referred to as "elliptic" or "primal" formulation.

As mentioned in the Introduction, typical examples of applications for Darcy's equation are oil reservoir engineering problems, as well as groundwater flow simulations for the management of drinking water reservoirs, see e.g. [23, 24, 35]. These are very complex problems, characterized by eterogeneous material properties and by large temporal and spatial scales. It is easy to realize that significant amount of uncertainty affects the majority of the parameters; as a consequence, the application of efficient Uncertainty Quantification techniques in this context is a very relevant research area.

The deterministic Darcy problem can be approximated numerically in its mixed form, see e.g. [13, 86] and references therein. Indeed, such methods have attracted a growing interest in the last decades, since they can efficiently and accurately handle problems in which the material characteristics feature large and sudden oscillations, and the numerical solutions they provide exhibit interesting conservation properties. Their application in a Stochastic Galerkin setting has been explored e.g. in [32]. Nonetheless, in this work we will consider a standard Finite Element discretization of the primal formulation (4.2) of the Darcy problem, and we will apply the error-work grids derived in Chapter 3.

The rest of this Chapter is organized as follows. In Section 4.2 we specify the modelistic assumptions on the random permeability field, on the deterministic problem and on the quantity of interest. Section 4.3 deals with the finite dimensional Fourier expansion of the random field, and Section 4.5 with the derivation of the optimal sparse grid for the problem at hand. Finally, we present some numerical results in Section 4.6, and draw some conclusions in Section 4.7.

4.2 Problem setting

We will consider the case in which the permeability field a is the only source of uncertainty for the problem. Let D be a bounded domain in $\mathbb{R}^d$ and $(\Omega, \mathcal{F}, P)$ be a complete probability space, where Ω denotes the set of outcomes, $\mathcal{F}$ its σ-algebra, and $P : \mathcal{F} \to [0,1]$ a probability measure. We assume that the permeability field can be modeled as a random field $a(\mathbf{x}, \omega)$, for which the following assumptions hold:

Assumption 4.1. *$a = a(\mathbf{x}, \omega) : \overline{D} \times \Omega \to \mathbb{R}$ is a random field on $(\Omega, \mathcal{F}, P)$, such that for $\mathbf{p}, \mathbf{q} \in D$, the covariance function $C_a(\mathbf{p}, \mathbf{q}) = \mathbb{C}\text{ov}\,[a(\mathbf{p}, \cdot)a(\mathbf{q}, \cdot)]$ depends only on the distance $\|\mathbf{p} - \mathbf{q}\|$ ("weak stationarity" property). Moreover, $C_a(\mathbf{p}, \mathbf{q}) = C_a(\|\mathbf{p} - \mathbf{q}\|)$ is Lipschitz continuous.*

Given the assumption of Lipschitz continuity of C_a, the Kolmogorov continuity theorem (see e.g. [80]) allows to conclude that the trajectories of a are a.s. continuous over $\overline{D}$. In particular, it is possible to prove (see [19]) the following proposition.

Proposition 4.1. *The random field a admits a version whose trajectories belong to the space of Hölder continuous functions $C^{0,\alpha}(\overline{D})$ a.s. with $\alpha < 1/2$.*

Following the notation of the previous chapters, we denote with $H^1(D)$ the Sobolev space of square-intergrable functions in D with square integrable derivatives, and its dual with $H^1(D)'$. $L^q_P(\Omega)$ will denote the Banach space of random functions with bounded q-th moment with respect to the probability measure P, and $L^q_P(\Omega; H^1(D))$ the Bochner space of $H^1(D)$-valued random fields with q-th bounded moment with respect to P, that is

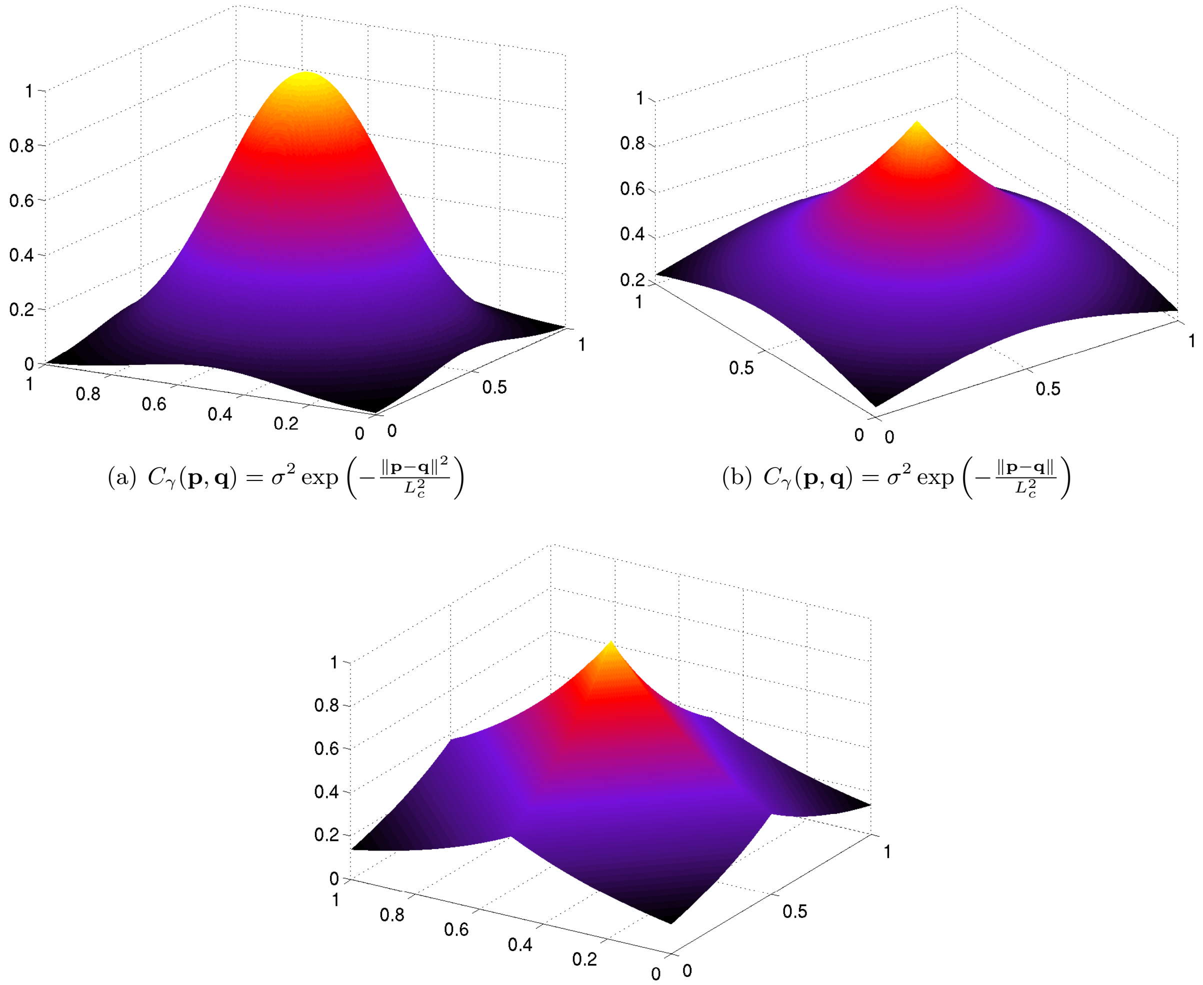
$$f \in L^q_P(\Omega; H^1(D)) \Leftrightarrow \int_\Omega \|f(\cdot, \omega)\|^q_{H^1(D)}\, dP(\omega) < \infty\,.$$

In addition to Assumption 4.1, we further need to choose the probability distribution for $a(\mathbf{x}, \cdot)$ and the functional shape of $C_a(\|\mathbf{p} - \mathbf{q}\|)$. Since hydrogeologycal applications deal in general with composite materials (sand, marl, clay), the pointwise permeability value can experience variations of orders of magnitude. It is more appropriate therefore to describe the logarithm of the permeability rather than the permability itself as a random field, and in particular we will make the following assumption, which is largely common in hydrogeologycal applications:

Assumption 4.2. *$a(\mathbf{x}, \omega)$ is a lognormal field, that is*

$$a(\mathbf{x}, \cdot) = e^{\gamma(\mathbf{x}, \cdot)}, \quad \gamma(\mathbf{x}, \cdot) \sim \mathcal{N}(\mu, \sigma^2) \quad \forall \mathbf{x} \in D, \tag{4.3}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian probability distribution with expected value μ and variance σ^2. Note that after such hypothesis, the random field is not uniformely bounded nor uniformely coercive with respect to ω. However, as shown in [19], since D is bounded and the trajectories of a are a.s. continuous on $\overline{D}$, there exist two real-valued random variables $a_{min}(\omega) = \min_{\mathbf{x} \in \overline{D}} a(\mathbf{x}, \omega)$, $a_{max}(\omega) = \max_{\mathbf{x} \in \overline{D}} a(\mathbf{x}, \omega)$, with the following properties:



(a) $C_\gamma(\mathbf{p},\mathbf{q}) = \sigma^2 \exp\left(-\frac{\|\mathbf{p}-\mathbf{q}\|^2}{L_c^2}\right)$

(b) $C_\gamma(\mathbf{p},\mathbf{q}) = \sigma^2 \exp\left(-\frac{\|\mathbf{p}-\mathbf{q}\|}{L_c^2}\right)$

(c) $C_\gamma(\mathbf{p},\mathbf{q}) = \sigma^2 \exp\left(-\frac{\|\mathbf{p}-\mathbf{q}\|_1}{L_c^2}\right)$, with $\|\mathbf{p}-\mathbf{q}\|_1 = |p_1-q_1| + |p_2-q_2|$

Figure 4.1: Examples of covariance function for an isotropic field. In the plots $\mathbf{p} = (0.5, 0.5)$ and $\mathbf{q} \in D = (0,1)^2$

Proposition 4.2. $1/a_{min}(\omega)$, $a_{max}(\omega) \in L^q_P(\Omega)$, $\forall q > 0$.

In particular, the summability of $a_{min}(\omega)$, $a_{max}(\omega)$ is a consequence of Fernique's theorem [33].

As for the covariance function, several models have been proposed in the literature. Such function is usually a decreasing function controlled by two parameters: the variance in each point σ^2 and the "correlation length" L_c. Common examples are shown in figure 4.1. While a somehow more realistic choice for hydrologycal application is to model the covariance function for the exponent field γ as an Exponential covariance function (fig. 4.1(b)-4.1(c)), it is intuitive that the spike featured by the Exponential covariance function will make the problem quite difficult to tackle. As a consequence, given the exploratory level of this work, we choose here to work with the more regular Gaussian covariance function (fig. 4.1(a)),

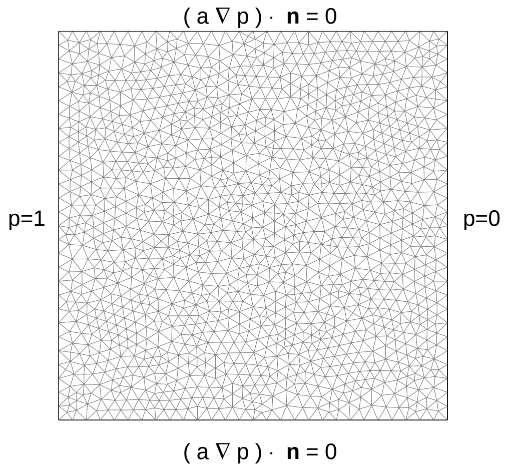


Figure 4.2: Computational domain for problem (4.5) with boundary conditions. The computational mesh used is also shown

Assumption 4.3. $\gamma(\mathbf{x},\omega)$ *has a Gaussian covariance function,*

$$C_\gamma(\mathbf{p},\mathbf{q}) = \sigma^2 \exp\left(-\frac{\|\mathbf{p}-\mathbf{q}\|^2}{L_c^2}\right). \tag{4.4}$$

The Darcy problem will be set in a horizontal square domain $D = (0,1)^2$ with no forcing terms. We define the following portions of the boundary ∂D:

$$\begin{aligned}\partial D_1 &= \{\,(x_1,x_2) : x_1 = 0\,\}\\ \partial D_2 &= \{\,(x_1,x_2) : x_1 = 1\,\}\\ \partial D_3 &= \{\,(x_1,x_2) : x_2 = 0\,\}\\ \partial D_4 &= \{\,(x_1,x_2) : x_2 = 1\,\},\end{aligned}$$

and let $\mathbf{n}$ be the outward normal to ∂D. We impose a pressure gradient acting on the water by setting $p = 1$ on the left boundary ∂D_1 and $p = 0$ on the right boundary ∂D_2. Finally, we consider a no-flux Neumann condition on the upper and lower boundaries ∂D_3 and ∂D_4, see Figure 4.2. The Darcy problem (4.2) thus reads:

Strong Formulation. *Find a random pressure* $p : \overline{D} \times \Omega \to \mathbb{R}$ *such that* P*-almost everywhere the following equation holds*

$$\begin{cases} -\operatorname{div}(a(\mathbf{x},\omega)\nabla p(\mathbf{x},\omega)) = 0 & \mathbf{x} \in D,\\ p(\mathbf{x},\omega) = 1 & \mathbf{x} \in \partial D_1,\\ p(\mathbf{x},\omega) = 0 & \mathbf{x} \in \partial D_2,\\ a(\mathbf{x},\omega)\nabla p(\mathbf{x},\omega)\cdot\mathbf{n} = 0 & \mathbf{x} \in \partial D_3 \cup \partial D_4, \end{cases} \tag{4.5}$$

It is straightforward to see that, thanks to the Lax–Milgram lemma, (4.5) is well-posed for almost every $\omega \in \Omega$, that is, for almost every $\omega \in \Omega$ there exists a unique pressure p solving (4.5), which can be understood as an $H^1(D)$-valued random field over $(\Omega, \mathcal{F}, P)$. Moreover, it holds

$$\|p(\cdot,\omega)\|^q_{H^1(D)} \leq \frac{1}{a^q_{min}(\omega)} \|f\|^q_{H^1(D)'}\,, \forall q > 0. \tag{4.6}$$

Proving the well-posedness of (4.5) in the Bochner spaces $L^q_P(\Omega; H^1(D))$ for $q > 0$ is then a consequence of Proposition 4.2, see again [19].

Proposition 4.3. *For every $q > 0$, there exists a unique $H^1(D)$-valued random pressure $p = p(\mathbf{x}, \omega)$ in $L^q_P(\Omega; H^1(D))$ solving* (4.5)*.*

The well-posedness of (4.5) has been proved also in [37, 43]. As for quantities of interest, we are interested in computing the expected value of the total flux crossing the right boundary ∂D_4. This is indeed a random variable,

$$Z_p(\omega) = \int_{\partial D_4} a(\mathbf{x}, \omega) \nabla p(\mathbf{x}, \omega) d\mathbf{x}\,, \tag{4.7}$$

and also represents the "effective permeability" of the random medium in D.

4.3 Series expansion of the log-permeability random field

Equation (4.3) represents $\gamma(\mathbf{x}, \omega)$ (and hence $a(\mathbf{x}, \omega)$) as a function of an infinite number of random variables, one per point $\mathbf{x} \in D$. To get to a computable representation of a we need therefore to derive an approximation of a in terms of a finite set of N random variables $y_i(\omega)$ ("finite noise approximation"). Such approximation can be obtained by means of suitable truncations of decompositions such as the Karhunen-Loève expansion, see [42, 62, 63, 64].

Proposition 4.4 (Karhunen-Loève expansion)**.** *Let $\gamma(\mathbf{x}, \omega)$ be a random field on $(\Omega, \mathcal{F}, P)$ with continuous covariance function $C_\gamma(\mathbf{p}, \mathbf{q})$. The operator T defined as*

$$v \in L^2(D) \to T(v) = \int_D C_\gamma(\mathbf{x}, \mathbf{x}') v(\mathbf{x}') d\mathbf{x}' \in L^2(D),$$

is a linear, symmetric and compact operator. Therefore, it admits a decreasing and non-negative sequence of eigenvalues $\{\lambda_i\}_{i \in \mathbb{N}}$, and the corresponding eigenvectors $\{v_i\}_{i \in \mathbb{N}}$ form an orthonormal basis for $L^2(D)$. Then it holds

$$\gamma(\mathbf{x}, \omega) = \mathbb{E}\left[\gamma(\mathbf{x}, \cdot)\right] + \sum_{i=1}^{\infty} \sqrt{\lambda_i} v_i(\mathbf{x}) y_i(\omega), \tag{4.8}$$

with $y_i(\omega)$ uncorrelated random variables with zero mean and unit variance, defined as

$$y_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_D \left(\gamma(\mathbf{x}, \omega) - \mathbb{E}\left[\gamma(\mathbf{x}, \cdot)\right] \right) v_i(\mathbf{x})\, d\mathbf{x}.$$

Moreover,

$$\int_D \mathbb{V}\text{ar}\left[\gamma(\mathbf{x}, \cdot)\right] d\mathbf{x} = \sum_{i=1}^{\infty} \lambda_i\,. \tag{4.9}$$

Observe that since we have assumed that $\gamma(\mathbf{x}, \cdot)$ is a Gaussian random variable for each $\mathbf{x} \in D$ (see Assumption 4.2), y_i are Gaussian too, and therefore independent. A finite noise approximation of γ can then be obtained simply retaining only the first N terms of (4.8),

$$\gamma_N(\mathbf{x}, \omega) = \mathbb{E}\left[\gamma(\mathbf{x}, \cdot)\right] + \sum_{i=1}^{N} \sqrt{\lambda_i} v_i(\mathbf{x}) y_i(\omega), \tag{4.10}$$

where N has to be large enough to include in the truncated expansion a sufficient percentage of the total variability of γ according to equation (4.9). The truncated expansion γ_N converges to γ thanks to the Mercer's theorem, and the convergence is driven by the decay of the eigenvalues of the covariance function, see [3, 101] and references therein. In particular, it holds

$$\sup_{\mathbf{x}\in D} \mathbb{E}\left[\left(\gamma(\mathbf{x},\cdot) - \gamma_N(\mathbf{x},\cdot)\right)^2\right] = \sup_{\mathbf{x}\in D} \sum_{i=N+1}^{\infty} \lambda_i v_i^2(\mathbf{x}) \to 0 \text{ when } N \to \infty\,.$$

The decay of λ_i is in turn related to the spatial regularity of the covariance function. Roughly speaking, the smoother the covariance, the faster the decay of λ. In particular, an analytic covariance function will result in an exponential decay of λ_i, hence in an exponential convergence of γ_N to γ, while a covariance function with finite Sobolev regularity will instead result in an algebraic decay only. Therefore, we expect that the Karhunen-Loève expansion of a field with a Gaussian covariance (fig. 4.1(a)) will require fewer terms than the expansion of a field with an Exponential covariance (fig. 4.1(b)-4.1(c)). The decay of λ_i is also influenced by the size of the correlation length: small correlation lengths indeed need a richer truncation (4.10) to be correctly representend.

In analogy with the properties of a spectral decomposition of a matrix, the Karhunen-Loève expansion is an optimal decomposition of a random field, in the sense that the truncated expansion (4.10) is the minimizer of the $L^2_P(\Omega)$ approximation error, i.e. among all the possible N-terms expansions of γ built upon an orthonormal basis for $L^2(D)$, the Karhunen-Loève expansion is the one that explains the highest part of the total variability of γ.

As an alternative to the Karhunen-Loève expansion, we consider in this work a Fourier-based decompostion of γ, which uses trigonometric polynomials as basis functions in the physical space, thus highlighting the contribution of each frequency to the total field a.

Proposition 4.5 (Fourier expansion). *Let $\gamma(\mathbf{x},\omega) : [0,L]^2 \times \Omega \to \mathbb{R}$ be a weakly stationary random field as in Assuption 4.1, with point-wise variance σ^2. The covariance function of $\gamma(\mathbf{x},\omega)$ can be expanded in cosine-Fourier series,*

$$C_\gamma(\|\mathbf{p}-\mathbf{q}\|) = \sigma^2 \sum_{\mathbf{k}=(k_1,k_2)\in\mathbb{N}_0^2} c_{\mathbf{k}} \cos(\omega_{k_1}(p_1-q_1))\cos(\omega_{k_2}(p_2-q_2)), \tag{4.11}$$

with $\omega_{k_1} = \frac{k_1\pi}{L}$, $\omega_{k_2} = \frac{k_2\pi}{L}$, and normalized coefficient $c_{\mathbf{k}}$ so that

$$\sum_{\mathbf{k}\in\mathbb{N}_0^2} c_{\mathbf{k}} = 1. \tag{4.12}$$

The random field admits then the following representation

$$\gamma(\mathbf{x},\omega) = \mathbb{E}\left[\gamma(\mathbf{x},\cdot)\right] + \sigma \sum_{\mathbf{k}\in\mathbb{N}_0^2} \sqrt{c_{\mathbf{k}}}\,[\, y^1_{\mathbf{k}}(\omega)\cos(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2) + y^2_{\mathbf{k}}(\omega)\sin(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2)$$
$$+y^3_{\mathbf{k}}(\omega)\cos(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2) + y^4_{\mathbf{k}}(\omega)\sin(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2)\,], \tag{4.13}$$

where $y^i_{\mathbf{k}}(\omega)$ are identically distributed and uncorrelated random variables with zero mean and unit variance.

See the Appendix of this Chapter for a proof of this Proposition. Again, since γ is a Gaussian random field, $y_i(\omega)$ are Gaussian random variables, hence independent. We can again define a truncated expansion of γ,

$$\gamma_N(\mathbf{x},\omega) = \mathbb{E}\left[\gamma(\mathbf{x},\cdot)\right] + \sigma \sum_{\mathbf{k}\in\mathcal{K}} \sqrt{c_{\mathbf{k}}}\,[\, y^1_{\mathbf{k}}(\omega)\cos(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2) + y^2_{\mathbf{k}}(\omega)\sin(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2)$$
$$+y^3_{\mathbf{k}}(\omega)\cos(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2) + y^4_{\mathbf{k}}(\omega)\sin(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2)\,], \tag{4.14}$$

where $\mathcal{K} \subset \mathbb{N}_0^2$ is a index set big enough to take into account a sufficient amount of the total variability of γ. Note that to compute the number of random variables included in γ_N by the truncation (4.14) one has to be keep into account that for $\mathbf{k} = (0,0)$ the only non-zero contribution is given by $y^1_{\mathbf{k}}(\omega)$, and similarly $\mathbf{k} = (0,k)$ and $\mathbf{k} = (k,0)$ will result in only two non-zero contributions.

The convergence of the Fourier expansion (4.14) is related to the decay of the coefficients $c_{\mathbf{k}}$ of the cosine-Fourier expansion of the covariance function. Similarly to the Karhunen-Loève expansion, a smoother covariance function will result in a faster decay of $c_{\mathbf{k}}$ and hence to a faster convergence.

Remark 4.1. *The computation of the Fourier series implies the periodic extension of the covariance function outside $[-L, L]$. The resulting periodic function may however feature low regularity at $x = \pm L$, that in turn would imply a slow convergence of the Fourier series. To prevent this, one may think of considering the Fourier series of the covariance periodically extended over a broader interval $[-\widetilde{L}, \widetilde{L}]$, with $\widetilde{L} > L$ and $\widetilde{L} \gg L_c$ large enough so that the covariance function in $\pm\widetilde{L}$ is almost zero. This approach has been adopted in this Chapter.*

Now let $\Gamma_i = \mathbb{R}$ denote the support of $y_i(\omega)$, $\Gamma = \Gamma_1 \times \ldots \times \Gamma_N$ the support of $\mathbf{y} = [y_1, \ldots, y_N]$, $\rho_i(y_i) : \Gamma_i \to \mathbb{R}$ the probability density function of y_i and $\rho(\mathbf{y}) : \Gamma \to \mathbb{R}$ the probability density function of $\mathbf{y}$, with

$$\rho(\mathbf{y}) = \prod_{n=1}^{N} \rho_i(y_i), \qquad \rho_i(y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y_i^2}{2}}.$$

Having introduced the random variables y_i, we can replace the abstract probability space $(\Omega, \mathcal{F}, P)$ with $(\Gamma, \mathcal{B}(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, where $\mathcal{B}(\Gamma)$ denotes the Borel σ-algebra, and hence $L^q_P(\Omega)$ with $L^q_\rho(\Gamma)$ and $L^q_P(\Omega; H^1(D))$ with $L^q_\rho(\Gamma; H^1(D))$. Moreover, the permeability and pressure fields can now be seen as functions of $\mathbf{x}$ and $\mathbf{y}$, so that $a(\mathbf{x}, \mathbf{y})$ can be approximated as $a(\mathbf{x}, \mathbf{y}) \approx a_N(\mathbf{x}, \mathbf{y}) = e^{\gamma_N(\mathbf{x}, \mathbf{y})}$, and problem (4.5) can be recast as

Strong Formulation. *Find a random pressure $p_N : \overline{D} \times \Gamma \to \mathbb{R}$ such that $\rho(\mathbf{y})d\mathbf{y}$-almost everywhere the following equation holds*

$$\begin{cases} -\operatorname{div}(a_N(\mathbf{x}, \mathbf{y}) \nabla p_N(\mathbf{x}, \mathbf{y})) = 0 & \mathbf{x} \in D, \\ p_N(\mathbf{x}, \mathbf{y}) = 1 & \mathbf{x} \in \partial D_1, \\ p_N(\mathbf{x}, \mathbf{y}) = 0 & \mathbf{x} \in \partial D_2, \\ a_N(\mathbf{x}, \mathbf{y}) \nabla p_N(\mathbf{x}, \mathbf{y}) \cdot \mathbf{n} = 0 & \mathbf{x} \in \partial D_3 \cup \partial D_4, \end{cases} \tag{4.15}$$

and the quantity of interest (4.7) becomes a random function $Z_p : \Gamma \to \mathbb{R}$,

$$Z_p(\mathbf{y}) = \int_{\partial D_4} a_N(\mathbf{x}, \mathbf{y}) \nabla p_N(\mathbf{x}, \mathbf{y}) d\mathbf{x}. \tag{4.16}$$

As the number of random variables N included in (4.10)-(4.14) increases, the truncated pressure p_N converges to p. Note that even if N is chosen such that the truncated log-permeability field γ_N includes a given percentage of the total variability of γ, say $\alpha\%$, this does not imply that a_N includes $\alpha\%$ of the total variability of a, and *a fortiori* p_N will not in general include $\alpha\%$ of the variability of p. However, the study on the convergence of p_N to p will not be addressed in this chapter, see e.g. [19] to this end. Therefore in the rest of this chapter, with a slight abuse of notation, we will remove the subscript and denote p_N with p. Moreover, the Optimal Sparse Grid Collocation technique that we will present in the next Section is able to automatically select the "most important" random variables that should be retained for the approximation of p. This would allow us to work with formally $N \to \infty$ random variables.

Similarly to the Strong Formulation (4.5) over the abstract probability space $(\Omega, \mathcal{F}, P)$, Lax–Milgram lemma ensures that for almost every $\mathbf{y} \in \Gamma$ there exists a unique $H^1(D)$-valued random pressure defined over $(\Gamma, \mathcal{B}(\Gamma), \rho(\mathbf{y})d\mathbf{y})$ solving (4.15), with

$$\|p(\cdot, \mathbf{y})\|_{H^1(D)} \leq \frac{1}{a_{N,min}(\mathbf{y})} \|f\|_{H^1(D)'} . \tag{4.17}$$

Then, using again Proposition 4.2, we obtain the well-posedness of (4.15) in $L^q_\rho(\Gamma; H^1(D))$, for all $q > 0$.

Proposition 4.6. *For every $q > 0$, there exists a unique $H^1(D)$-valued random pressure $p = p(\mathbf{x}, \mathbf{y})$ in $L^q_\rho(\Gamma; H^1(D))$ solving* (4.15).

In particular, denoting with $H^1_{dir}(D)$ the subset of $H^1(D)$ functions that vanish on the Dirichlet boundary $\partial D_1 \cup \partial D_2$, the following Weak Formulation is well-posed:

Weak Formulation. *Find $p \in H^1(D) \otimes L^2_\rho(\Gamma)$ such that $p = 1$ on ∂D_1, $p = 0$ on ∂D_2 and $\forall\, v \in H^1_{dir}(D) \otimes L^2_\rho(\Gamma)$*

$$\int_\Gamma \int_D a_N(\mathbf{x}, \mathbf{y}) \nabla p(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y})\, \rho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = 0. \tag{4.18}$$

Note that in such weak formulation we look for a pressure with a separable structure for $\mathbf{x}$ and $\mathbf{y}$. This is motivated by the fact that the Karhunen-Loève and Fourier expansions (4.10)-(4.14) used to derive (4.15) also assume such separable structure for a_N.

4.4 Auxiliary norms

Proposition 4.6 ensures the well-posedness of (4.15) in all the Bochner spaces $L^q_\rho(\Gamma; H^1(D))$ with $q > 0$. If on the one hand these are the most natural spaces for the problem at hand, they are on the other hand unsatisfactory spaces to a certain extent, since the usual inclusion $L^\infty_\rho(\Gamma) \subset L^q_\rho(\Gamma)$, with

$$L^\infty_\rho(\Gamma) = \left\{ f : \Gamma \to \mathbb{R} : \operatorname*{ess\,sup}_{\mathbf{y} \in \Gamma} |f(\mathbf{y})\rho(\mathbf{y})| < \infty \right\}, \quad \|f\|_{L^\infty_\rho(\Gamma)} = \operatorname*{ess\,sup}_{\mathbf{y} \in \Gamma} |f(\mathbf{y})\rho(\mathbf{y})| ,$$

does not hold true. Equivalently, the norms $\|\cdot\|_{L^q_\rho(\Gamma)}$, $\|\cdot\|_{L^\infty_\rho(\Gamma)}$ do not satisfy $\|\cdot\|_{L^q_\rho(\Gamma)} \leq C \|\cdot\|_{L^\infty_\rho(\Gamma)}$, Indeed, such inequality holds only if $0 < q < 1$, since

$$\begin{aligned} \int_\Gamma f^q(\mathbf{y}) e^{-\frac{\sum_n y_n^2}{2}} d\mathbf{y} &= \int_\Gamma f^q(\mathbf{y}) \left(e^{-\frac{\sum_n y_n^2}{2}} \right)^q \left(e^{-\frac{\sum_n y_n^2}{2}} \right)^{-q} e^{-\frac{\sum_n y_n^2}{2}} d\mathbf{y} \\ &\leq \|f\|^q_{L^\infty_\rho(\Gamma)} \int_\Gamma \left(e^{-\frac{\sum_n y_n^2}{2}} \right)^{-q} e^{-\frac{\sum_n y_n^2}{2}} d\mathbf{y} = \|f\|^q_{L^\infty_\rho(\Gamma)} \int_\Gamma e^{-(1-q)\frac{\sum_n y_n^2}{2}} d\mathbf{y}. \end{aligned}$$

It may be therefore of interest to introduce an auxiliary measure $\widetilde{\rho}(\mathbf{y})d\mathbf{y}$ such that both $\|\cdot\|_{L^q_\rho(\Gamma)} \leq C \|\cdot\|_{L^\infty_{\widetilde{\rho}}(\Gamma)}$ and $p \in L^\infty_{\widetilde{\rho}}(\Gamma)$ are satisfied. To this end, the following lemma holds.

Lemma 4.1. *Given problem (4.15), with $a_N = e^{\gamma_N(\mathbf{x},\mathbf{y})}$ and $\gamma_N(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^N \beta_n \phi_n(\mathbf{x}) y_n$, the measure*

$$\widetilde{\rho}(\mathbf{y}) = e^{-\sum_{n=1}^N \alpha_n |y_n|}, \quad \alpha_n = \beta_n \|\phi_n\|_{L^\infty(D)} , \tag{4.19}$$

is such that

(i) $\|\cdot\|_{L^q_\rho(\Gamma)} \leq \widetilde{C}(N,\alpha,q)\,\|\cdot\|_{L^\infty_{\widetilde{\rho}}(\Gamma)}$*, where* $\widetilde{C}(N,\alpha,q) = \prod_{n=1}^N e^{\alpha_n^2 q/2}\sqrt[q]{2\Phi(\alpha_n q)}$ *and* $\Phi(t)$ *is the cumulative distribution function of a standard Gaussian distribution.*

(ii) $p \in L^\infty_{\widetilde{\rho}}(\Gamma)$*, with* $\|p\|_{L^\infty_{\widetilde{\rho}}(\Gamma)} \leq \|f\|_{H^1(D)'}$.

Moreover, if $\sum_{n=1}^\infty \alpha_n < +\infty$ *then it holds* $\lim_{N\to\infty} \widetilde{C}(N,\alpha,q) < +\infty$.

Proof. (i) holds since

$$\begin{aligned}
\int_\Gamma f^q(\mathbf{y}) \frac{e^{-\frac{\sum_n y_n^2}{2}}}{(2\pi)^{N/2}}\,d\mathbf{y} &= \int_\Gamma f^q(\mathbf{y}) \left(e^{-\sum_n \alpha_n |y_n|}\right)^{-q} \left(e^{-\sum_n \alpha_n |y_n|}\right)^{q} \frac{e^{-\frac{\sum_n y_n^2}{2}}}{(2\pi)^{N/2}}\,d\mathbf{y} \\
&\leq \|f\|^q_{L^\infty_{\widetilde{\rho}}(\Gamma)} \int_\Gamma e^{q\sum_n \alpha_n |y_n|} \frac{e^{-\frac{\sum_n y_n^2}{2}}}{(2\pi)^{N/2}} d\mathbf{y} = \|f\|^q_{L^\infty_{\widetilde{\rho}}(\Gamma)} \prod_{n=1}^N \int_{-\infty}^{\infty} e^{q\alpha_n|y_n|} \frac{e^{-\frac{y_n^2}{2}}}{\sqrt{2\pi}} dy_n \\
&= \|f\|^q_{L^\infty_{\widetilde{\rho}}(\Gamma)} \prod_{n=1}^N e^{\alpha_n^2 q^2/2} \left[\int_{-\infty}^0 \frac{e^{-\frac{(y_n+\alpha_n q)^2}{2}}}{\sqrt{2\pi}} dy_n + \int_0^\infty \frac{e^{-\frac{(y_n-\alpha_n q)^2}{2}}}{\sqrt{2\pi}} dy_n \right] \\
&= \|f\|^q_{L^\infty_{\widetilde{\rho}}(\Gamma)} \prod_{n=1}^N 2 e^{\alpha_n^2 q^2/2} \Phi(\alpha_n q).
\end{aligned}$$

As for (ii), using the Lax–Milgram estimate (4.17) and the expression of γ_N we obtain

$$\begin{aligned}
\|p\|_{L^\infty_{\widetilde{\rho}}(\Gamma)} &= \sup_{\mathbf{y}\in\Gamma} \left| \|p(\cdot,\mathbf{y})\|_{H^1(D)}\, \widetilde{\rho}(\mathbf{y}) \right| \leq \sup_{\mathbf{y}\in\Gamma} \left| \frac{1}{a_{N,min}(\mathbf{y})} \|f\|_{H^1(D)'}\, e^{-\sum_n \alpha_n |y_n|} \right| \\
&\leq \sup_{\mathbf{y}\in\Gamma} \left| e^{-\min_{\mathbf{x}\in D}\left(\sum_n \beta_n \phi_n(\mathbf{x}) y_n\right)} \|f\|_{H^1(D)'}\, e^{-\sum_n \alpha_n |y_n|} \right| \\
&\leq \|f\|_{H^1(D)'} \sup_{\mathbf{y}\in\Gamma} \left| e^{\sum_n \left(\beta_n \|\phi_n\|_{L^\infty(D)} - \alpha_n\right)|y_n|} \right|,
\end{aligned}$$

and the proof concludes choosing $\alpha_n = \beta_n \|\phi_n\|_{L^\infty(D)}$, with $\beta_n = \sqrt{\lambda_n}$ for a Karhunen-Loève expansion of γ and $\beta_n = c_n$ for a Fourier expansion.

Finally, to compute $\lim_{N\to\infty} \widetilde{C}(N,\alpha,q)$ we first introduce the auxiliary quantity $v(N,\alpha,q)$ as

$$v(N,\alpha,q) = \log \widetilde{C}(N,\alpha,q) = \sum_{n=1}^N \left(\alpha_n^2 \frac{q}{2} + \frac{1}{q} \log\left(2\Phi(\alpha_n q)\right) \right).$$

Next we exploit the fact that for $x > 0$ it holds that $\Phi(x) \leq \frac{1}{2} + \frac{x}{\sqrt{2\pi}}$, to obtain

$$\log\left(2\Phi(x)\right) \leq \log\left(1 + \sqrt{\frac{2}{\pi}}x\right) \leq \sqrt{\frac{2}{\pi}}x,$$

and therefore

$$v(N,\alpha,q) \leq \sum_{n=1}^N \left(\alpha_n^2 \frac{q}{2} + \frac{1}{q}\sqrt{\frac{2}{\pi}}\alpha_n q \right).$$

Now, let $S_1 = \sum_{n=1}^\infty \alpha_n$ and $S_2 = \sum_{n=1}^\infty \alpha_n^2$. With this notation, we have $v(N,\alpha,q) \leq \frac{q}{2}S_2 + \sqrt{\frac{2}{\pi}}S_1$, therefore we can compute

$$\lim_{N\to\infty} \widetilde{C}(N,\alpha,q) = \lim_{N\to\infty} e^{v(N,\alpha,q)} \leq e^{\frac{q}{2}S_2 + \sqrt{\frac{2}{\pi}}S_1},$$

which is finite provided $S_1, S_2 < \infty$. □

Remark 4.2. *The condition $\sum_{n=1}^{\infty} \alpha_n < \infty$ is not always satisfied for a Karhunen-Loève/Fourier expansion, and has to be explicitly verified for each case.*

Remark 4.3. *The measure $\widetilde{\rho}(\mathbf{y})d\mathbf{y}$ in general will not be a probability measure, since $\int_\Gamma \widetilde{\rho}(\mathbf{y})d\mathbf{y} = \int_\Gamma e^{-\sum_{n=1}^{N}\alpha_n |y_n|}d\mathbf{y} = \prod_{n=1}^{N}\frac{2}{\alpha_n} \neq 1$.*

4.5 Optimal sparse grid approximation

An accurate representation of γ through a truncated Karhunen-Loève or Fourier expansion (4.10)-(4.14) will in general depend on a high number of of random variables y_i, as the following Example shows.

Example 4.1. *For sufficiently small values of L_c the coefficients of the cosine-Fourier transform* (4.11) *of the Gaussian covariance function* (4.4) *are well approximated by*

$$c_{\mathbf{k}} \approx \lambda_{k_1}\lambda_{k_2}, \quad \lambda_k = \begin{cases} \dfrac{L_c\sqrt{\pi}}{2L} & \text{if } k = 0 \\ \dfrac{L_c\sqrt{\pi}}{L}\exp\left(-\dfrac{(k\pi L_c)^2}{4L^2}\right) & \text{if } k > 0\ . \end{cases} \tag{4.20}$$

see the Appendix of this chapter for a proof. An efficient truncation (4.14) *of the Fourier expansion of $\gamma(\mathbf{x}, \omega)$ should include all the harmonics $\mathbf{k} \in \mathcal{K}$ such that $c_{\mathbf{k}}$ is greater than a given threshold value. Since each coefficient $c_{\mathbf{k}}$ in* (4.20) *is proportional to $k_1^2 + k_2^2$, we consider here the truncation defined by*

$$\mathcal{K} = \left\{\mathbf{k} : k_1^2 + k_2^2 \leq w,\ w \in \mathbb{N}\right\}. \tag{4.21}$$

Note that, thanks to (4.12)*, if $\sum_{\mathbf{k}\in\mathcal{K}} c_{\mathbf{k}} = \alpha$ then γ_N is taking into account $\alpha\%$ of the total variability of the field. Table 4.1 shows the number of random variables that need to be included into the series* (4.14) *to take into account $\alpha\%$ of the total variability of γ for different correlation lengths L_c. The need to include a high number of random variables in the approximation of the random field γ, and hence the high-dimensionality of p, clearly emerges.*

	$\boldsymbol{\alpha = 0.7}$	$\boldsymbol{\alpha = 0.9}$	$\boldsymbol{\alpha = 0.99}$
$\mathbf{L_c = 0.35}$	$N = 13$	$N = 25$	$N = 49$
$\mathbf{L_c = 0.25}$	$N = 25$	$N = 49$	$N = 97$
$\mathbf{L_c = 0.1}$	$N = 161$	$N = 293$	$N = 593$

Table 4.1: Random variables needed to represent $\alpha\%$ of the total variability of a random field with Gaussian covariance function for different correlation lengths L_c.

To handle the high-dimensionality of p efficiently, we now derive the optimal sparse grid approximation of $p(\mathbf{x}, \mathbf{y})$, extending the procedure presented in Chapter 3 to the lognormal case considered.

4.5.a Abstract construction of optimal sparse grids

Recall that the sparse grid approximation of p is defined as

$$p_w(\mathbf{y}) = \mathcal{S}^m_{\mathcal{I}(w)}[p](\mathbf{y}) = \sum_{\mathbf{i}\in\mathcal{I}(w)} \bigotimes_{n=1}^{N} \Delta_n^{m(i_n)}[p](\mathbf{y}), \tag{4.22}$$

where $\{\mathcal{I}(w)\}_{w\in\mathbb{N}}$ denotes a sequence of index sets, $\Delta^{m(\mathbf{i})}[p] = \bigotimes_{n=1}^{N} \Delta^{m(i_n)}[p]$ is called hierarchical surplus and $\Delta_n^{m(i)} = \mathcal{U}_n^{m(i)} - \mathcal{U}_n^{m(i-1)}$ is the difference between two consecutive one-dimensional interpolants over $m(i)$ and $m(i-1)$ points respectively. In Chapter 3 we have detailed an a-priori procedure to derive optimal sparse grids, based on estimates of the profit of each hierarchical surplus, i.e. using

$$\mathcal{I}(w) = \left\{ \mathbf{i} \in \mathbb{N}_+^N : \frac{\Delta E(\mathbf{i})}{\Delta W(\mathbf{i})} \geq \epsilon(w) \right\}^{ADM} \tag{4.23}$$

in (4.22), where $\{\epsilon(w)\}_{w\in\mathbb{N}} \downarrow 0$ and $\Delta E(\mathbf{i})$, $\Delta W(\mathbf{i})$ represent the error and work contribution of each hierarchical surplus, see Chapter 3 for details.

In particular, if the considered interpolant operators $\mathcal{U}_n^{m(i_n)}$ are nested and the set $\mathcal{I}(w)$ is admissible, the work contribution can be computed exactly as

$$\Delta W(\mathbf{i}) = \prod_{n=1}^{N} (m(i_n) - m(i_n - 1)). \tag{4.24}$$

Before stating the error contribution estimate, we need to introduce a spectral basis for $L^2_\rho(\Gamma)$. To this end, let $\{H_p(y_n)\}_{p\in\mathbb{N}}$ be the family of orthonormal Hermite polynomials relative to the weight $e^{-y^2/2}/\sqrt{2\pi}$ in the n-th direction,

$$\int_{\Gamma_n} H_p(y_n) H_q(y_n) \rho_n(y_n) dy_n = \delta_{p,q}, \quad \forall p, q \in \mathbb{N},$$

that can be computed either recursively (see [39]) or through the explicit formula

$$H_q(y_n) = \frac{(-1)^q}{\sqrt{q!}} e^{y_n^2/2} \frac{d^q}{dy_n^q} e^{-y_n^2/2}.$$

It is then immediate to see that the set of multidimensional Hermite polynomials

$$\mathcal{H}_{\mathbf{q}}(\mathbf{y}) = \prod_{n=1}^{N} H_{q_n}(y_n), \quad \forall \mathbf{q} \in \mathbb{N}^N \tag{4.25}$$

is an orthonormal basis for $L^2_\rho(\Gamma)$, that can be used to construct the spectral expansion of $p(\mathbf{y})$

$$p(\mathbf{y}) = \sum_{\mathbf{q}\in\mathbb{N}^N} p_{\mathbf{q}} \mathcal{H}_{\mathbf{q}}(\mathbf{y}), \quad p_{\mathbf{q}} = \int_\Gamma p(\mathbf{y}) \mathcal{H}_{\mathbf{q}}(\mathbf{y}) \rho(\mathbf{y}) d\mathbf{y}. \tag{4.26}$$

We are now in position to state an heuristic estimate for the error contribution of the hierarchical surplus $\Delta^{m(\mathbf{i})}$ in the spirit of what was done in Chapter 3, equation (3.33):

$$\Delta E(\mathbf{i}) \approx B(\mathbf{i}) \left\| p_{m(\mathbf{i-1})} \right\|_{H^1(D)}, \tag{4.27}$$

where $p_{m(\mathbf{i-1})}$ is the $m(\mathbf{i-1})$-th coefficient of the spectral expansion (4.26), and $B(\mathbf{i})$ is a factor that in Chapter 3 was chosen equal to the product of the Lebesgue constant for the interpolant operator $\mathcal{U}_n^{m(i)}$, $B(\mathbf{i}) = \prod_{n=1}^{N} \mathbb{L}_n^{m(i_n)}$, based on numerical experiments. To conclude the optimal sparse grid construction we thus need to choose a family of nested interpolant operators for the Gaussian measure $\rho(\mathbf{y})$, estimate the factor $B(\mathbf{i})$ (with an heuristic argument) and the spectral coefficient $p_{m(\mathbf{i}-1)}$.

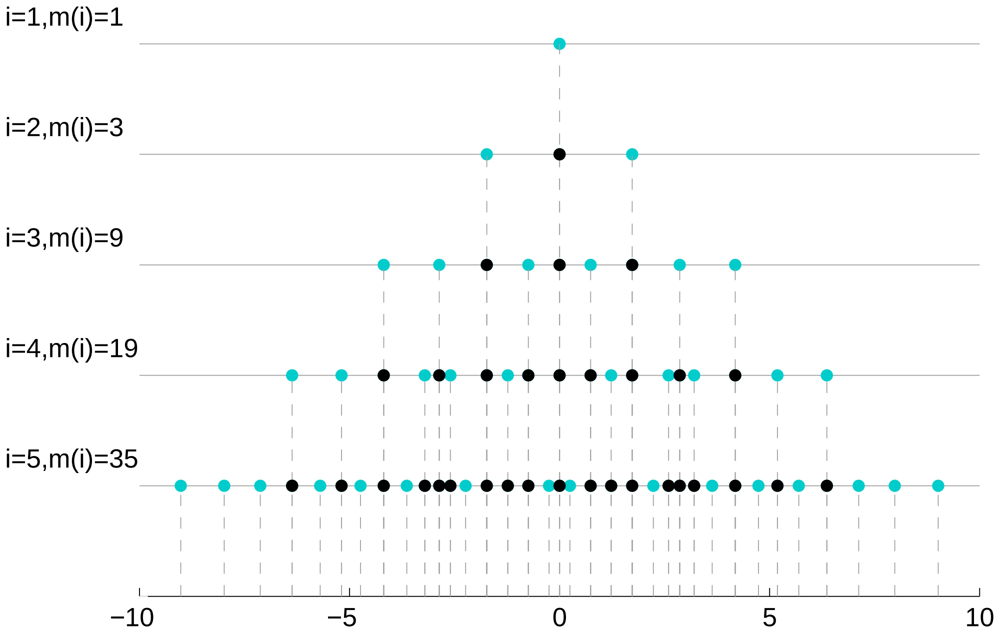


Figure 4.3: *KPN* knots for levels going from $i = 1$ to $i = 5$. The new points added at each level are marked in light blue. For each level i we also show the points of the corresponding standard Gauss–Hermite quadrature formula with $m(i)$ knots (gray crosses).

4.5.b Nested quadrature formulae for Gaussian measure

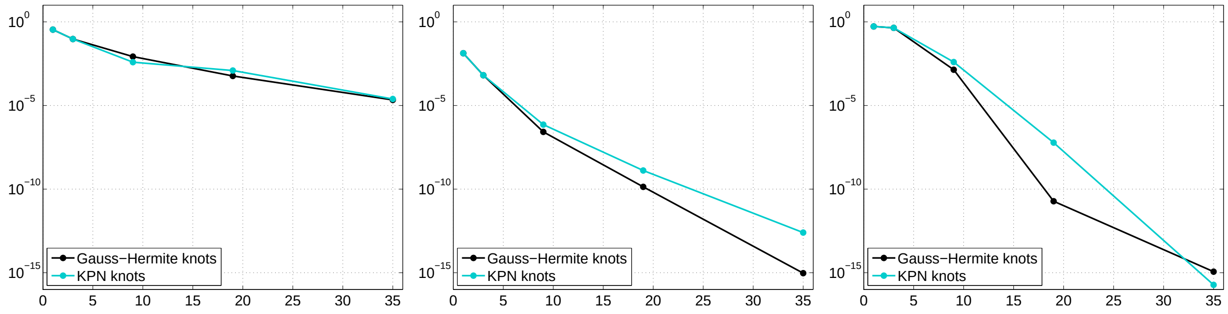
The family of nested points we choose to use in this chapter are the so-called "Kronrod-Patterson-Normal" nested points (*KPN* in short). Such family of interpolation/quadrature points is due to Genz and Keister, see [40, 48], that applied the Kronrod-Patterson procedure [58, 81] to the classical Gauss-Hermite quadrature points (i.e. the roots of the Hermite polynomials $H_p(y_n)$). We recall that the Kronrod-Patterson procedure is a way to modify a quadrature rule, by adding new points in a nested fashion retaining the highest accuracy possible. Details are given in the Appendix of this Chapter.

Let $\{y_{i,k}^{KPN}\}_{k=1,\dots,m(i)}$ denote the set of *KPN* points at the i-th interpolation level. The knots and the corresponding quadrature weights are tabulated up to level 5 (35 nodes) and can be found e.g. at `http://www.sparse-grids.de/`. Figure 4.3 shows the *KPN* knots for levels $i = 1, \dots, 5$; note that the function $m(i)$ does not have a regular pattern.

Figure 4.4 compares the error convergence of *KPN* quadrature with the standard Gauss–Hermite quadrature. The *KPN* rule is thus seen to have good performances, close to the Gauss–Hermite ones.

4.5.c Estimate for $B(\mathbf{i})$

In Chapter 3 the constant $B(\mathbf{i})$ in equation (4.27) was chosen to be equal to the product of the Lebesgue constants of interpolant operators in each direction, $B(\mathbf{i}) = \prod_{n=1}^{N} \mathbb{L}_n^{m(i_n)}$. This is a reasonable heuristic assumption, since in this way the error contribution estimate "encodes" information on both the quality of the solution (through the decay of the spectral coefficients), and the quality of the interpolant operator itself. Such an estimate is also supported by numerical validation.



(a) $f(t) = 1/(1+t^2)$ (b) $f(t) = 1/(2+e^t)$ (c) $f(t) = \cos(3t+1)$

Figure 4.4: comparison of performance of *KPN* and Gauss–Hermite quadrature rules for the approximation of $1/\sqrt{2\pi}\int_{-\infty}^{\infty} f(t)e^{-t^2/2}dt$.

To extend the choice $B(\mathbf{i}) = \prod_{n=1}^{N} \mathbb{L}_n^{m(i_n)}$ to the lognormal case, we need to suitably extend the definition of the Lebesgue constant to the problem at hand. Yet, the natural norm $L^\infty_\rho(\Gamma)$ does not appear a suitable norm to be used in this context, since we have seen that it does not satisfy the usual inequality $\|\cdot\|_{L^q_\rho(\Gamma)} \leq C\,\|\cdot\|_{L^\infty_\rho(\Gamma)}$. Using the auxiliary norms in Lemma 4.1 to define the Lebesgue constant one gets

$$\mathbb{L}_n^{m(i)} = \sup_{v\in C^0_{\widetilde{\rho}}(\Gamma_n)} \frac{\left\|\mathcal{U}_n^{m(i)}v\right\|_{L^\infty_{\widetilde{\rho}}(\Gamma_n)}}{\|v\|_{L^\infty_{\widetilde{\rho}}(\Gamma_n)}} = \sup_{v\in C^0_{\widetilde{\rho}}(\Gamma_n)} \frac{\sup_{y_n\in\Gamma_n}\left|\widetilde{\rho}(y_n)\sum_{k=0}^{m(i)} l_{k,n}^{m(i)}(y_n)v(y_{n,k})\right|}{\sup_{y_n\in\Gamma_n}|v(y_n)\widetilde{\rho}(y_n)|}$$

where $l_{k,n}^{m(i)}(y_n)$, $k = 1,\dots,m(i)$ denotes the Lagrangian polynomials associated to the *KPN* knots at the i-th interpolation level. However, it is not easy to obtain a sharp bound for such quantity.

Thus we propose a numerical estimate for $B(\mathbf{i})$, which gives good numerical results when tested on model problems (see Figure 4.7) and at the same time is close to the original choice $B(\mathbf{i}) = \prod_{n=1}^{N}\mathbb{L}_n^{m(i_n)}$ when applied to a problem with uniform random variables.

To this end, we go back to the definition of error contribution for a hierarchical surplus, and exploit the fact that p admits an Hermite expansion:

$$\begin{aligned}\Delta E(\mathbf{i}) &= \left\|\left(p - \mathcal{S}^m_{\{\mathcal{J}\cup\mathbf{i}\}}[p]\right) - \left(p - \mathcal{S}^m_{\mathcal{J}}[p]\right)\right\|_{H^1(D)\otimes L^2_\rho(\Gamma)} = \left\|\Delta^{m(\mathbf{i})}[p]\right\|_{H^1(D)\otimes L^2_\rho(\Gamma)} \\ &= \left\|\Delta^{m(\mathbf{i})}\Big[\sum_{\mathbf{q}\in\mathbb{N}^N} p_{\mathbf{q}}\mathcal{H}_{\mathbf{q}}\Big]\right\|_{H^1(D)\otimes L^2_\rho(\Gamma)} = \left\|\sum_{\mathbf{q}\in\mathbb{N}^N} p_{\mathbf{q}}\Delta^{m(\mathbf{i})}[\mathcal{H}_{\mathbf{q}}]\right\|_{H^1(D)\otimes L^2_\rho(\Gamma)}.\end{aligned} \tag{4.28}$$

Next, since $\Delta^{m(\mathbf{i})}$ is a difference of two interpolant operators, the lower order Hermite polynomials will be interpolated exactly by both interpolants and their contribution to the summation will be zero. Then, by triangular inequality we get to

$$\Delta E(\mathbf{i}) = \left\|\sum_{\mathbf{q}\geq m(\mathbf{i}-\mathbf{1})} p_{\mathbf{q}}\Delta^{m(\mathbf{i})}[\mathcal{H}_{\mathbf{q}}]\right\|_{H^1(D)\otimes L^2_\rho(\Gamma)} \leq \sum_{\mathbf{q}\geq m(\mathbf{i}-\mathbf{1})} \|p_{\mathbf{q}}\|_{H^1(D)}\left\|\Delta^{m(\mathbf{i})}[\mathcal{H}_{\mathbf{q}}]\right\|_{L^2_\rho(\Gamma)}. \tag{4.29}$$

Therefore, the error estimate (4.27) is equivalent to assuming that the summation at the right-hand

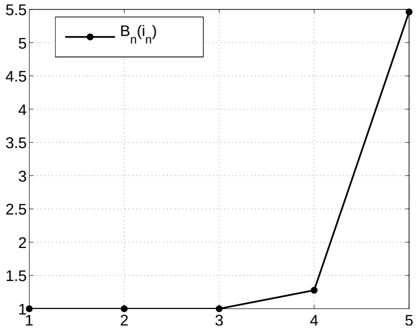


Figure 4.5: Numerical results for the computation of $B_n(i_n) = \left\|\mathcal{U}^{m(i_n)}[H_{m(i_n-1)}(y_n)] - \mathcal{U}^{m(i_n-1)}[H_{m(i_n-1)}(y_n)]\right\|_{L^2_{\rho_n}(\Gamma_n)}$. The integrals have been computed exactly with a Gauss–Hermite quadrature formula with a sufficient number of quadrature points.

side of (4.29) is dominated by the first term, with

$$B(\mathbf{i}) = \left\|\Delta^{m(\mathbf{i})}[\mathcal{H}_{m(\mathbf{i-1})}]\right\|_{L^2_\rho(\Gamma)} = \prod_{n=1}^{N} B_n(i_n)\,, \tag{4.30}$$

$$B_n(i_n) = \left\|\Delta^{m(i_n)}[H_{m(i_n)}]\right\|_{L^2_{\rho_n}(\Gamma_n)} = \left\|\mathcal{U}^{m(i_n)}[H_{m(i_n-1)}(y_n)] - \mathcal{U}^{m(i_n-1)}[H_{m(i_n-1)}(y_n)]\right\|_{L^2_{\rho_n}(\Gamma_n)}\,.$$

The quantity $B_n(i_n)$ can be easily computed numerically, and results in a moderate growth with respect to i_n, see Figure 4.5.

Remark 4.4. *The procedure used here to derive an estimate for $B(\mathbf{i})$ could be applied to the problems investigated in Chapter 3, where we have considered uniform random variables rather than Gaussian ones. The result would be $B(\mathbf{i}) = \prod_n B_n(i_n)$, $B_n(i_n) = \left\|\Delta^{m(i_n)}[L_{m(i_n-1)}(y_n)]\right\|_{L^2_{\mathcal{U}_n}(\Gamma_n)}$, where $\Gamma_n = [-1,1]$, $L^2_{\mathcal{U}_n}(\Gamma_n)$ is the space of the random functions square integrable with respect to the uniform weight $\rho_n(y_n) = 1/2$ and $L_n(y_n)$ denotes the corresponding orthonormal Legendre polynomials. However, considering $B_n(i_n)$ as the Lebesgue constant of the Clenshaw–Curtis points, $B_n(i_n) = \mathbb{L}^{m(i_n)}_{CC}$, instead of $B_n(i_n) = \left\|\Delta^{m(i_n)}[L_{m(i_n-1)}(y_n)]\right\|_{L^2_{\mathcal{U}}(\Gamma_n)}$, does not affect significantly the results, since the two quantities are close, at least for values of i_n used in practical cases, as shown in Figure 4.6.*

Next, we test such choice on the model function $p(y_1, y_2) = 1/\exp(1 + b_1 y_1 + b_2 y_2)$, so that we can compute each $\Delta E(\mathbf{i})$ as $\Delta E(\mathbf{i}) = \left\|\Delta^{m(\mathbf{i})}[p]\right\|_{L^2_\rho(\Gamma)}$ using a sufficiently accurate sparse grid quadrature, see equation (4.28). The Hermite coefficients of p can be computed either numerically or analytically, see Lemma 4.2 in the next section. Once such quantities are available, we can verify the accuracy of (4.27), with $B(\mathbf{i})$ as in (4.30). The results are shown in Figure 4.7: the proposed estimate is thus seen to be quite reasonable.

4.5.d Convergence of Hermite expansions

Finally, we need to estimate the decay of the norm of the coefficients of the spectral expansion (4.26) of p, that will be used in (4.27) together with the estimate of $B(\mathbf{i})$ derived in the previous section.

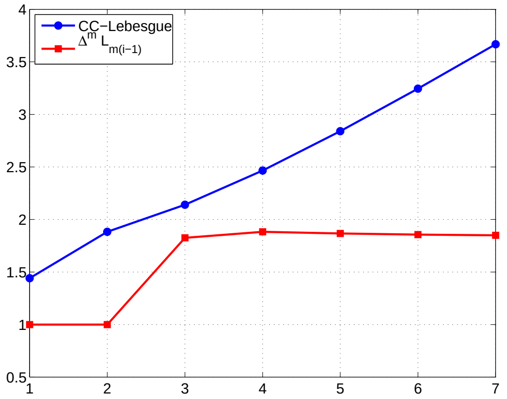
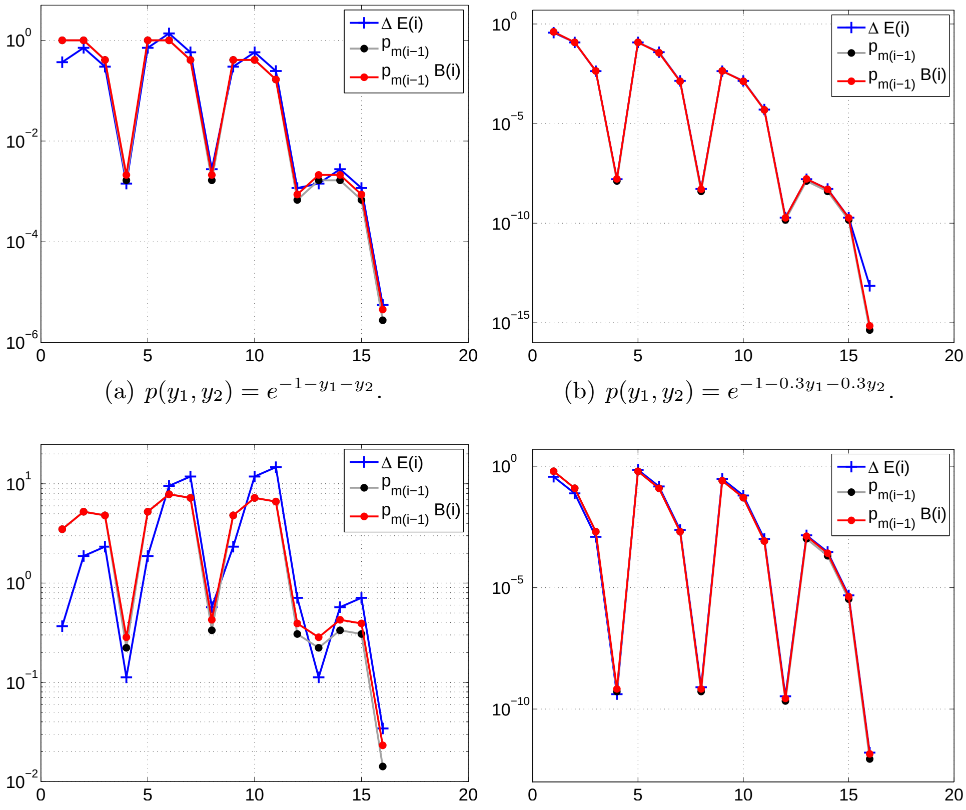


Figure 4.6: Comparison between $\left\|\Delta^{m(i)}[\mathcal{L}_{m(i-1)}]\right\|_{L^2_{\mathcal{U}}(\Gamma)}$ and $\mathbb{L}_{CC}^{m(i)}$.



(a) $p(y_1, y_2) = e^{-1-y_1-y_2}$.

(b) $p(y_1, y_2) = e^{-1-0.3y_1-0.3y_2}$.

(c) $p(y_1, y_2) = e^{-1-1.5y_1-1.5y_2}$.

(d) $p(y_1, y_2) = e^{-1-y_1-0.2y_2}$.

Figure 4.7: Numerical comparison between $\Delta E(\mathbf{i})$ and $|p_{m(\mathbf{i-1})}|$ for p of the form $p(y_1, y_2) = e^{-1-b_1y-1-b_2y_2}$. The quantities $\Delta E(\mathbf{i})$ for $\mathbf{i} \in TP(4)$ have been computed with a standard Smolyak sparse grid, with $\mathcal{I}(w) = \{\mathbf{i} \in \mathbb{N}^N_+ : |\mathbf{i} - \mathbf{1}| \leq w\}$, $w = 10$, and "doubling" function $m(i)$: $m(0) = 0, m(1) = 1, m(i) = 2^{i-1} + 1$. The Hermite coefficients $|p_{m(\mathbf{i-1})}|$ have been computed analytically with the formula stated in Lemma 4.2.

To derive an estimate for $\|p_{\mathbf{q}}\|_{H^1(D)}$ we first consider a simplified Darcy problem with a lognormal permeability field a constant over D, $a = a(\mathbf{y}) = \exp\left(b_0 + \sum_{i=1}^N b_i y_i\right)$ and with homogeneous Dirichlet boundary conditions,

$$\begin{cases} -\operatorname{div}\left(a(\mathbf{y})\nabla p(\mathbf{x},\mathbf{y})\right) = f(\mathbf{x}) & \mathbf{x} \in D, \\ p(\mathbf{x},\mathbf{y}) = 0 & \mathbf{x} \in \partial D. \end{cases} \tag{4.31}$$

Denoting with $h \in H^1(D)$ the solution of the auxiliary deterministic problem

$$\begin{cases} \Delta h(\mathbf{x}) = f(\mathbf{x}) & \mathbf{x} \in D, \\ h(\mathbf{x}) = 0 & \mathbf{x} \in \partial D, \end{cases} \tag{4.32}$$

we can write the analytic expression for p solving (4.31), which is separable with respect to $\mathbf{y}$,

$$p(\mathbf{x},\mathbf{y}) = h(\mathbf{x}) e^{-b_0} \prod_{n=1}^N \exp\left(-\, b_n y_n\right), \tag{4.33}$$

Thus, given the separable form of the multidimensional Hermite polynomials $\mathcal{H}_{\mathbf{p}}$ in eq. (4.25), it is enough to estimate the decay of the coefficients of the Hermite expansion of $\exp(-b_n y_n)$ to derive an estimate for the decay of $\|p_{\mathbf{q}}\|_{H^1(D)}$. To this end, the following lemma holds:

Lemma 4.2. *Given problem* (4.31)*, the* $H^1(D)$ *norm of the Hermite coefficients* (4.26) *of* p *can be bounded as*

$$\|p_{\mathbf{q}}\|_{H^1(D)} \leq C_{\mathcal{H}} \prod_{n=1}^N \frac{e^{-g_n q_n}}{\sqrt{q_n!}}, \tag{4.34}$$

with $C_{\mathcal{H}} = \|h\|_{H^1(D)} e^{-b_0} \prod_{n=1}^N e^{b_n^2/2}$ *and* $g_n = -\log(b_n)$*.*

Proof. Let us consider the Hermite expansion of $v_n(y_n) = e^{-b_n y_n}$,

$$v_n(y_n) = \sum_{i=0}^{\infty} v_{n,i} H_i(y_n), \quad v_{n,i} = \int_{\Gamma_n} v_n(y_n) H_i(y_n) \rho_n(y_n) dy_n.$$

Using (4.33) and (4.25) in (4.26), we have that $\|p_{\mathbf{q}}\|_{H^1(D)} = \|h\|_{H^1(D)} e^{-b_0} \prod_{n=1}^N |v_{n,q_n}|$, and we only need to estimate $|v_{n,q_n}|$. It holds

$$\begin{aligned} v_{n,q_n} &= \int_{\Gamma_n} v_n(y_n) H_{q_n}(y_n) \rho_n(y_n) dy_n = \int_{\Gamma_n} e^{-b_n y_n} \frac{(-1)^{q_n}}{\sqrt{q_n!}} e^{y_n^2/2} \frac{d^{q_n}}{dy_n^{q_n}} \left(e^{-y_n^2/2}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{y_n^2}{2}} dy_n \\ &= (-1)^{q_n} \frac{1}{\sqrt{2\pi q_n!}} \int_{\Gamma_n} e^{-b_n y_n} \frac{d^{q_n}}{dy_n^{q_n}} e^{-y_n^2/2} dy_n = (-1)^{2q_n} \frac{1}{\sqrt{2\pi q_n!}} \int_{\Gamma_n} \left(\frac{d^{q_n}}{dy_n^{q_n}} e^{-b_n y_n} \right) e^{-y_n^2/2} dy_n \\ &= (-1)^{3q_n} \frac{b_n^{q_n}}{\sqrt{2\pi q_n!}} \int_{\Gamma_n} e^{-b_n y_n} e^{-y_n^2/2} dy_n = (-1)^{3q_n} \frac{b_n^{q_n}}{\sqrt{2\pi q_n!}} e^{b_n^2/2} \int_{\Gamma_n} e^{-(y_n+b_n)^2/2} dy_n \\ &= (-1)^{3q_n} \frac{b_n^{q_n}}{\sqrt{q_n!}} e^{b_n^2/2}, \end{aligned}$$

and the proof is concluded by rewriting $b_n^{q_n} = e^{-g_n q_n}$ with $g_n = -\log(b_n)$. □

We now conjecture that estimate (4.34) is valid even in the more general case where $a_N(\mathbf{x},\mathbf{y}) = e^{\gamma_N(\mathbf{x},\mathbf{y})}$, and the boundary conditions are those specified in eq. (4.15). As pointed out in Chapter

3, in a general case the rates g_n are better estimated numerically. This is achieved by freezing all the variables y_i but the n^*-th one e.g. at the midpoint of thier support, and computing the solution $p_w^{n^*}$ of such reduced problem increasing the sparse grid level w from 1 to i^*. If the quadrature points are accurate enough (i.e. Gaussian quadrature points), then the intermediate solutions $p_w^{n^*}$ will converge to $p_{i^*}^{n^*}$ with the same trend as the spectral approximation, and the same holds for any quantity of interest $Z_p = Z_p(\mathbf{y})$ depending on p_w, that is

$$\left\| p_w^{n^*} - p_{i^*}^{n^*} \right\|_{L^2_\rho(\Gamma)\otimes H^1(D)} \leq C\, \frac{e^{-g_n m(w)}}{\sqrt{m(w)!}}, \quad \left\| Z_{p,w}^{n^*} - Z_{p,i^*}^{n^*} \right\|_{L^2_\rho(\Gamma)} \leq C\, \frac{e^{-g_n m(w)}}{\sqrt{m(w)!}}\,. \tag{4.35}$$

It is then possible to use a least square approach on the computed errors to derive an estimated value for g_n. Figure 4.8 shows the results of such procedure applied to a test case, and confirms the quality of the method proposed. Alternative estimates for the decay of the Hermite coefficients are available in [49].

Having estimated the Lebesgue constant of the *KPN* knots and the decay of the spectral coefficients $\|p_\mathbf{q}\|_{H^1(D)}$ we now have a computable expression for the work and error contributions (4.24)-(4.27), hence for the optimal set defintion (4.23). Thus, we obtain for the optimal set the following expression

$$\mathcal{I}^* = \left\{ \mathbf{i} \in \mathbb{N}_+^N : \frac{\prod_{n=1}^N B_n(i_n) \dfrac{e^{-g_n m(i_n-1)}}{\sqrt{m(i_n-1)!}}}{\prod_{n=1}^N (m(i_n) - m(i_n-1))} \geq \epsilon \right\}^{ADM},$$

that can be rewritten as

$$\mathcal{I}^* = \left\{ \mathbf{i} \in \mathbb{N}^N : \sum_{n=1}^N \left[g_n m(i_n-1) + \frac{1}{2}\log\big(m(i_n-1)!\big) - \log B_n(i_n) + \log\big(m(i_n) - m(i_n-1)\big) \right] \leq w \right\}^{ADM}, \tag{4.36}$$

where $w = \lceil -\log \epsilon \rceil \in \mathbb{N}$.

4.6 Numerical results

We are now in position to compute the optimal sparse approximation of p solving (4.15) and hence of the quantity of interest (4.16).

We consider the case of a stratified material in the direction transversal to the flow: that is, the log-permeability field γ depends only on x_1 and is constant along x_2. Thus the covariance function is

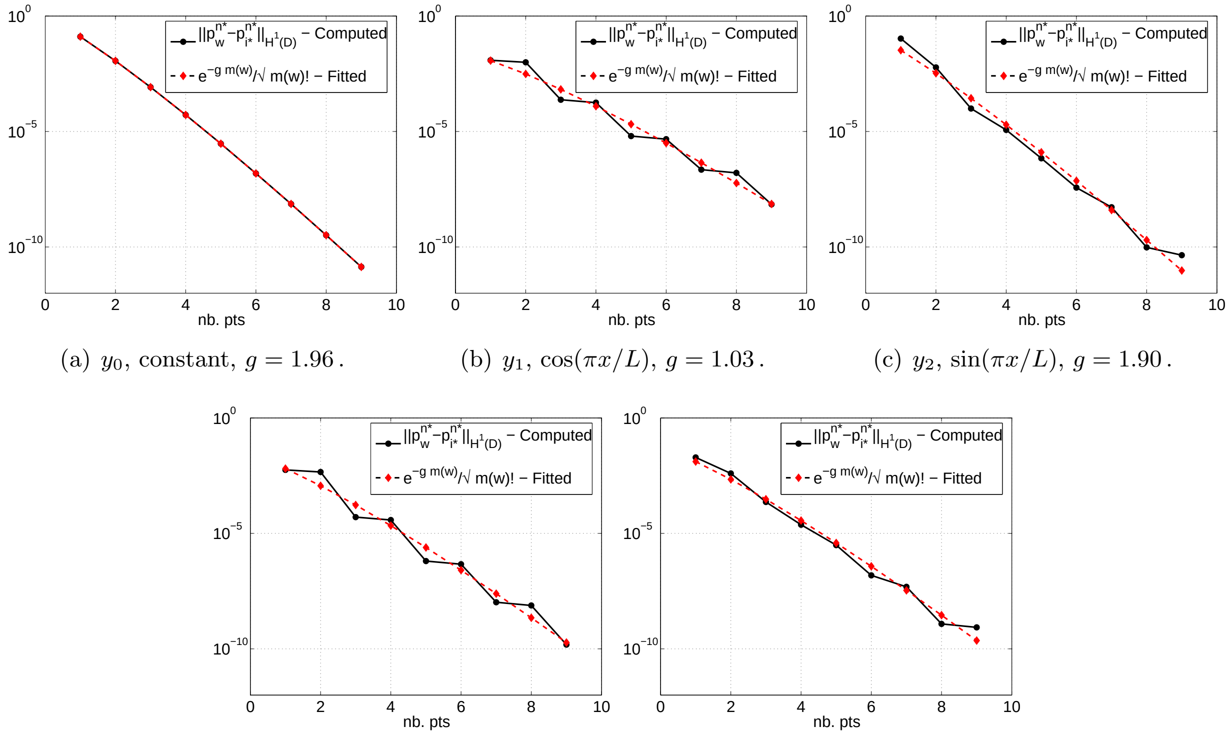
$$C_\gamma(s,t) = \sigma^2 \exp\left(-\frac{|s-t|^2}{L_c^2} \right), \quad s,t \in [0,1],$$

and the truncated Fourier expansion of γ_N (4.14) simplifies to

$$\gamma_N(x_1, \mathbf{y}) = \mathbb{E}\left[\gamma(\mathbf{x},\cdot)\right] + \sigma\sqrt{c_0}y_0 + \sigma \sum_{k=1}^K \sqrt{c_k}\,[\, y_{2k-1}\cos(\omega_k x_1) + y_{2k}\sin(\omega_k x_1)], \tag{4.37}$$

with $y_k \sim \mathcal{N}(0,1)$, $\omega_k = k\pi/L$, $L = 1$ and

$$c_k = \begin{cases} \dfrac{L_c\sqrt{\pi}}{2L} & \text{if } k = 0 \\ \dfrac{L_c\sqrt{\pi}}{L} \exp\left(-\dfrac{(k\pi L_c)}{2L} \right) & \text{if } k > 0\,. \end{cases} \tag{4.38}$$



(a) y_0, constant, $g = 1.96$.

(b) y_1, $\cos(\pi x/L)$, $g = 1.03$.

(c) y_2, $\sin(\pi x/L)$, $g = 1.90$.

(d) y_5, $\cos(3\pi x/L)$, $g = 1.39$.

(e) y_6, $\sin(3\pi x/L)$, $g = 1.31$.

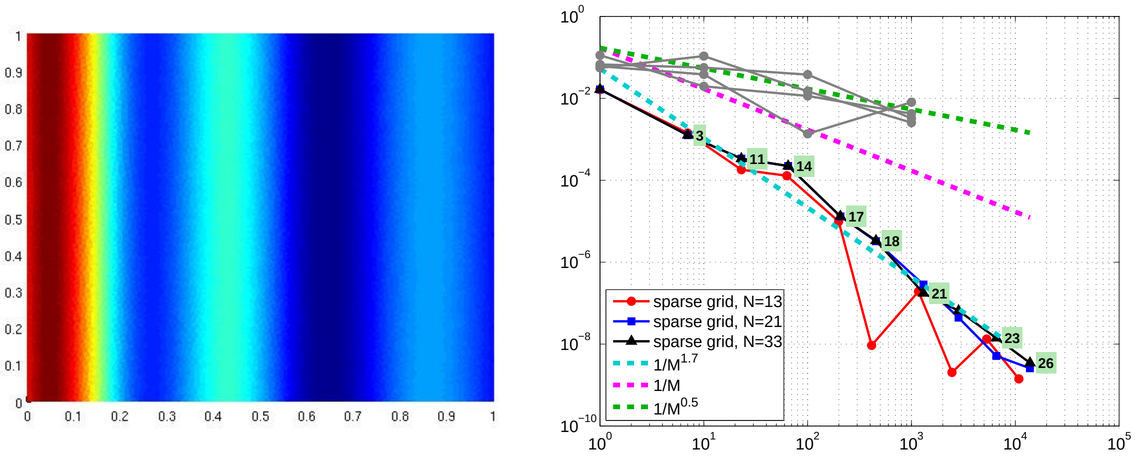
Figure 4.8: Assessment of the rates g_n, $n = 0, 1, 2, 5, 6$, used to build the optimal set (4.36), estimated according to equation (4.35). For each random variable y_n the corresponding harmonic in the Fourier expansion (4.37) is specified. The plots show the decay of $\left\| Z^{n^*}_{p,w} - Z^{n^*}_{p,i^*} \right\|_{H^1(D)}$ as a function of the number of point $m(w)$ (see eq. (4.16) for the definition of Z_p), its fitting according to the proposed estimate $\dfrac{e^{-g_n m(w)}}{\sqrt{m(w)}}$ and the resulting value of g_n.

The proofs of (4.37) and (4.38) are similar to their bidimensional analogous (4.13) and (4.20) stated in Proposition 4.5 and Example 4.1 respectively, see Appendix. We set the correlation length to $L_c = 0.2$ and the pointwise standard deviation to $\sigma = 0.3$; Figure 4.9(a) shows a realization of (4.37).

We consider three different levels of truncation for γ_N in (4.37): $K = 6, 10, 16$ corresponding to $N = 13, 21, 33$ random variables. With these truncation we take into account 99%,99.99% and 99.9999999999% respectively of the total variability of γ. For each truncation we compute the optimal sparse grid approximation p_w using the sets (4.36), and then compute the expected value for the total outgoing flux Z defined in (4.16), using the resulting sparse grid quadrature rule.

For each truncation level the approximation error is computed as $|\mathbb{E}[Z_{p_w}] - \mathbb{E}[Z_{p_*}]|$, with p_* reference solution computed with a highly refined sparse grid for the same truncation. This means that we will indeed have three reference solutions, since the interest of this numerical test is to monitor the convergence rate of the optimal sparse grid for a fixed truncation rather than looking for the level at which the convergence of a under-resolved truncation stagnates. We also perform a classical Monte Carlo simulation, repeated three times.

Results are shown in Figure 4.9(b). The Monte Carlo simulations converge with the expected



(a) Realization of (4.37).

(b) Convergence for MC and sparse grid methods. The numbers on the plot denote the number of random variables activated up to that level.

Figure 4.9: Numerical results for the test case presented.

rate 1/2; we also show the convergence rate 1 that would be obtained with a quasi-Monte Carlo method, like Sobol' sequences (see e.g. [16, 52, 70, 94]). As for the sparse grids approximation, it is important to observe that not only they all converge with a rate higher than 1/2, but such rate seems to be independent of the truncation level. This would mean that the strategy detailed in Section 4.5 is quite effective in reducing the deterioration of the performance of the standard sparse grids as the number of random variables increases. Indeed, the selection of the most profitable hierarchical surpluses manages to "activate " (i.e. to put interpolation points) only in those directions that are most useful in explaining the total variability of the solution, so that the less influent random variables get activated only when the approximation error is sufficiently low. The number shown on the convergence plot indicates the number of random variables activated up to that point.

4.7 Conclusions

In this Chapter we have considered a Darcy problem with uncertain permeability, modeled as a lognormal random field with Gaussian covariance function. We have then applied the optimal sparse grid paradigm derived in Chapter 3 to the problem at hand: to this end, we have introduced a nested quadrature/interpolation rule, and we have estimated numerically its Lebesgue constant. Finally, we have derived an estimate for the decay of the coefficients of the Hermite expansion of p.

We have applied the optimal sparse grid thus obtained to a test case describing a stratified material, that has been discretized with a Fourier expansion with $N = 13, 21$ and 33 random variables. Numerical results on this preliminary numerical test seem to suggest that the optimal sparse grid procedure achieves a convergence rate higher than the ones of the most common sampling methods. Moreover, it is quite effective in reducing considerably the degradation of the performance suffered by the standard sparse grids approach when the number of random variables increases.

Appendix

Proof of Proposition 4.5

Proposition 4.5. *Let $\gamma(\mathbf{x},\omega)$ be an isotropic random field with point-wise variance σ^2. The covariance function of $\gamma(\mathbf{x},\omega)$ can be expanded in cosine-Fourier series,*

$$C_\gamma(\|\mathbf{p}-\mathbf{q}\|) = \sigma^2 \sum_{\mathbf{k}=(k_1,k_2)\in\mathbb{N}_0^2} c_{\mathbf{k}} \cos(\omega_{k_1}(p_1-q_1))\cos(\omega_{k_2}(p_2-q_2)), \tag{4.39}$$

with normalized coefficient $c_{\mathbf{k}}$ so that

$$\sum_{\mathbf{k}\in\mathbb{N}_0^2} c_{\mathbf{k}} = 1. \tag{4.40}$$

The random field admits then the following representation

$$\begin{aligned}\gamma(\mathbf{x},\omega) = \mathbb{E}\left[\gamma(\mathbf{x},\cdot)\right] + \sigma \sum_{\mathbf{k}\in\mathbb{N}_0^2} \sqrt{c_{\mathbf{k}}}\,[\; & y^1_{\mathbf{k}}(\omega)\cos(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2) + y^2_{\mathbf{k}}(\omega)\sin(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2) \\ & + y^3_{\mathbf{k}}(\omega)\cos(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2) + y^4_{\mathbf{k}}(\omega)\sin(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2)\,],\end{aligned} \tag{4.41}$$

where $\omega_{k_1} = \frac{k_1\pi}{L}$, $\omega_{k_2} = \frac{k_2\pi}{L}$ and $y^i_{\mathbf{k}}(\omega)$ are identically distributed and uncorrelated random variables with zero mean and unit variance.

Proof. Since the field is stationary and isotropic, the covariance function depends only on $\|\mathbf{p}-\mathbf{q}\|$, hence is an even function and admits a cosine-Fourier series. The coefficient $c_{\mathbf{k}}$ of the Fourier series (4.39) are then normalized to put σ^2 in evidence.

Let us now rewrite (4.41) in a more convenient way, introducing the compact notation

$$\phi^i_{\mathbf{k}}(x_1,x_2) = \begin{cases} \cos(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2) & \text{if } i=1 \\ \sin(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2) & \text{if } i=2 \\ \cos(\omega_{k_1}x_1)\sin(\omega_{k_2}x_2) & \text{if } i=3 \\ \sin(\omega_{k_1}x_1)\cos(\omega_{k_2}x_2) & \text{if } i=4\ . \end{cases}$$

Let moreover $\bar\gamma = \mathbb{E}\left[\gamma(\mathbf{x},\cdot)\right]$. With this notation, the field expansion (4.41) reads

$$\gamma(\mathbf{x},\omega) = \bar\gamma + \sigma \sum_{\mathbf{k},i} \sqrt{c_{\mathbf{k}}}\, y^i_{\mathbf{k}}(\omega)\phi^i_{\mathbf{k}}(x_1,x_2).$$

Inserting this expansion into the definition of covariance, one gets

$$\begin{aligned} C_\gamma(\mathbf{p},\mathbf{q}) =& \mathbb{E}\left[\,(\,\gamma(\mathbf{p},\omega)-\bar\gamma)\,(\,\gamma(\mathbf{q},\omega)-\bar\gamma)\,\right] \\ =& \mathbb{E}\left[\sigma\sum_{\mathbf{k},i}\sqrt{c_{\mathbf{k}}}y^i_{\mathbf{k}}(\omega)\phi^i_{\mathbf{k}}(p_1,p_2)\sigma\sum_{\mathbf{l},j}\sqrt{c_{\mathbf{l}}}y^j_{\mathbf{l}}(\omega)\phi^j_{\mathbf{l}}(q_1,q_2)\right] \\ =& \sigma^2\sum_{\mathbf{k},i}\sum_{\mathbf{l},j}\sqrt{c_{\mathbf{k}}}\sqrt{c_{\mathbf{l}}}\phi^i_{\mathbf{k}}(p_1,p_2)\phi^j_{\mathbf{l}}(q_1,q_2)\int_\Omega y^i_{\mathbf{k}}(\omega)y^j_{\mathbf{l}}(\omega)dP(\omega). \end{aligned}$$

Next one exploites the fact that $y^i_{\mathbf{k}}(\omega)$ are mutually uncorrelated and with unit variance to simplify the previous sum to

$$C_\gamma(\mathbf{p},\mathbf{q}) = \sigma^2\sum_{\mathbf{k},i} c_{\mathbf{k}}\phi^i_{\mathbf{k}}(p_1,p_2)\phi^i_{\mathbf{k}}(q_1,q_2)\,. \tag{4.42}$$

Now we expand back the terms $\phi^i_{\mathbf{k}}(p_1,p_2), \phi^i_{\mathbf{k}}(q_1,q_2)$. The generic term in (4.42) becomes

$$\begin{aligned}c_{\mathbf{k}}[\,&\cos(\omega_{k_1}p_1)\cos(\omega_{k_2}p_2)\cos(\omega_{k_1}q_1)\cos(\omega_{k_2}q_2)+\sin(\omega_{k_1}p_1)\sin(\omega_{k_2}p_2)\sin(\omega_{k_1}q_1)\sin(\omega_{k_2}q_2)\\ &+\cos(\omega_{k_1}p_1)\sin(\omega_{k_2}p_2)\cos(\omega_{k_1}q_1)\sin(\omega_{k_2}q_2)+\sin(\omega_{k_1}p_1)\cos(\omega_{k_2}p_2)\sin(\omega_{k_1}q_1)\cos(\omega_{k_2}q_2)\,],\end{aligned}$$

that can be rewritten as

$$c_{\mathbf{k}}\big[\cos(\omega_{k_1}p_1)\cos(\omega_{k_1}q_1)+\sin(\omega_{k_1}p_1)\sin(\omega_{k_1}q_1)\big]\big[\cos(\omega_{k_2}p_2)\cos(\omega_{k_2}q_2)+\sin(\omega_{k_2}p_2)\sin(\omega_{k_2}q_2)\big],$$

so that (4.42) is equivalent to

$$C_\gamma(\mathbf{p},\mathbf{q})=\sigma^2\sum_{\mathbf{k}}c_{\mathbf{k}}\cos\big(\omega_{k_1}(p_1-q_1)\big)\cos\big(\omega_{k_2}(p_2-q_2)\big)$$

using standard trigonometric equalities. The cosine-Fourier transform of the covariance function has been recovered and the proof is completed. □

Proof of Example 4.1

Lemma 4.3. *For sufficiently small values of L_c,*[1] *the coefficients of the cosine-Fourier transform* (4.11) *of the Gaussian covariance function* (4.4) *are well estimated by*

$$c_{\mathbf{k}}\approx\lambda_{k_1}\lambda_{k_2},\quad \lambda_k=\begin{cases}\dfrac{L_c\sqrt{\pi}}{2L} & \text{if } k=0\\[2ex] \dfrac{L_c\sqrt{\pi}}{L}\exp\left(-\dfrac{(k\pi L_c)^2}{4L^2}\right) & \text{if } k>0\ .\end{cases}$$

Proof. $C_\gamma(\mathbf{p},\mathbf{q})$ can be rewritten as a product

$$C_\gamma(\mathbf{p},\mathbf{q})=\sigma\exp\left(-\frac{|p_1-q_1|^2}{L_c^2}\right)\sigma\exp\left(-\frac{|p_2-q_2|^2}{L_c^2}\right),$$

therefore its Fourier series can in turn be recast as a product

$$C_\gamma(\mathbf{p},\mathbf{q})=\sigma^2\sum_{k_1\in\mathbb{N}_0}\lambda_{k_1}\cos(\omega_{k_1}(p_1-q_1))\sum_{k_2\in\mathbb{N}_0}\lambda_{k_2}\cos(\omega_{k_2}(p_2-q_2)),$$

each factor being the cosine-Fourier series of $\exp\left(-\dfrac{|x|^2}{L_c^2}\right)$. To obtain an analytic expression for λ_k we resort to the exponential form of the Fourier series and to tabulated closed-form Fourier transforms.

On the one hand, the exponential form of the Fourier series of the restriction of a function $f:\mathbb{R}\to\mathbb{R}$ to $[-L,L]$ reads

$$f(x)=\sum_{k\in\mathbb{Z}}f_k e^{i\omega_k x},\quad f_k=\frac{1}{2L}\int_{-L}^{L}f(x)e^{-i\omega_k x}dx$$

and the coefficient λ_k of the trigonometric Fourier series is related to f_k as

$$\begin{aligned}\lambda_k&=f_k+f_{-k}\ \text{ if } k\geq 1\\ \lambda_0&=f_0\,.\end{aligned}\tag{4.43}$$

[1] numerically assessed bound: $L_c<0.35L$, where L is the length of the domain.

On the other hand, the Fourier transform of f reads

$$\mathcal{F}[f](\xi) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi\xi x} dx,$$

and if the support of f is included in $[-L, L]$ the Fourier transform evaluated at integer values of ξ gives the coefficients of the exponential form of the Fourier series:

$$f_k = \frac{1}{2L}\int_{-L}^{L} f(x) e^{-i\frac{\pi k}{L}x} dx = \frac{1}{2L}\int_{-\infty}^{\infty} f(x) e^{-i\frac{\pi k}{L}x} dx = \int_{-\infty}^{\infty} f(2zL) e^{-i2\pi kz} dz = \mathcal{F}[f(2zL)](k).$$

We can use this equality to derive an approximated analytic expression for the coefficients f_k of the exponenstial Fourier series for $f(x) = \exp\left(-\frac{|x|^2}{L_c^2}\right)$, provided L_c is sufficiently small. The Fourier transform of $e^{-\alpha x^2}$, $\alpha \in \mathbb{R}$ has a closed-form expression:

$$\mathcal{F}\left[e^{-\alpha x^2}\right](\xi) = \sqrt{\frac{\pi}{\alpha}} \exp\left(\frac{\pi^2 \xi^2}{\alpha}\right),$$

therefore

$$f_k \simeq \mathcal{F}[f(2zL)](k) = \mathcal{F}\left[\exp\left(-\frac{4|z|^2 L^2}{L_c^2}\right)\right](k) = \sqrt{\pi}\frac{L_c}{2L} \exp\left(-\frac{(\pi k L_c)^2}{4L^2}\right),$$

which also implies $f_{-k} = f_k$ (the Fourier transform of an even and real function is even and real). The proof is concluded using equality (4.43).

□

Deriving the *KPN* knots with the Kronrod-Patterson procedure

The Kronrod–Patterson procedure aims at adding points to a given quadrature rule in order to get the highest possible accuracy (see [58, 81]). In this Appendix, we apply such procedure to derive the *KPN* nested family of quadrature rules for the approximation of integrals with respect to a Gaussian density function as done in [40], which we will follow closely in the forthcoming exposition.

Background and notation

For $f \in \mathcal{C}^0(\mathbb{R})$, let $G(f)$ denote the integral of f over $[a, b]$ with respect to the density function $w(t)$, and $Q_N(f)$ a N-points quadrature rule for the numerical approximation of $G(f)$,

$$G(f) = \int_a^b f(t) w(t) dt \approx Q_N(f) = \sum_{i=1}^{N} f(t_j)\beta_j, \qquad t_j \in [a, b], \quad \beta_j \in \mathbb{R}.$$

A superscript D denotes the fact that Q_N^D has degree of exactness D, i.e. Q integrates exactly all polynomials with degree up to D, $Q^D(p) = G(p)$, $\forall p \in \mathbb{P}^D(\mathbb{R})$. Let also $\pi_N(t)$ denote the nodal polynomial associated to Q_N^D,

$$\pi_N(t) = \prod_{j=1}^{N} (t - t_j), \qquad \pi_N \in \mathbb{P}^N(\mathbb{R}). \tag{4.44}$$

Finally, since we are interested in iterative enrichments of a quadrature rule Q_N^D, it will be convenient to write $Q_N^D[i_1 + i_2 + \ldots + i_k]$ to denote a quadrature rule that has been obtained by the successive addition of i_1 points as a first step, i_2 as second step and so on.

Next, we state a fundamental result in integration theory, that will be used extensively througout this Section, see e.g. [85] for a proof.

Theorem 4.1. *For a given $m > 0$, the quadrature rule Q_N^D has degree of exactness $D = N + m - 1$ if and only if it is interpolatory and the nodal polynomial π_N is such that*

$$\int_a^b \pi_N(t)p(t)w(t)dt = 0 \qquad \forall p \in \mathbb{P}^{m-1}(\mathbb{R}). \tag{4.45}$$

Corollary 4.1. *The maximum degree of exactness for a quadrature rule with N points is $D = 2N - 1$.*

Proof. Observe that $m - 1$ must necessarily be lower than N, otherwise it would be possible to choose $p = \pi_N$ in (4.45). This would imply $\int_a^b \pi_N^2(t)w(t)dt = 0$, which is impossible since $w > 0$. Thus the maximum degree is $D = 2N - 1$. □

A quadrature rule that achieves $D = 2N - 1$ is called Gaussian quadrature rule. Equation (4.45) implies the well-known fact that the nodal polynomial of an N-point Gaussian quadrature rule with respect to the density w is the $(N + 1)$-th w-orthogonal polynomial.

In our case, $a = -\infty, b = \infty$ and $w(t) = e^{-t^2/2}/\sqrt{2\pi}$ is a symmetric density. The first quadrature rule of the *KPN* family is the one-point Gauss–Hermite quadrature rule Q_1^1, with $t_1 = 0$.

We require the subsequent quadrature rules to be nested, so that $t_1 = 0$ will be included in each rule. Moreover, since w is symmetric, each rule will be also symmetric. This can be obtained by adding at each step of the procedure couples of opposite points, $\pm\lambda_1, \pm\lambda_2, \ldots$, with equal weights for $t = \lambda_k$ and $t = -\lambda_k$, hence $N = 2\mu + 1$, for $\mu \in \mathbb{N}$.

The numbers λ_k are called generators, and we introduce the following short-hand notation for the evaluation of Q_N^D:

$$Q_N^D(f) = w_0 f(0) + w_1 f[\lambda_1] + w_2 f[\lambda_2] + \ldots + w_\mu f[\lambda_\mu], \tag{4.46}$$

with $f[\lambda_k] = \big(f(\lambda_k) + f(-\lambda_k)\big)$.

The optimal addition of quadrature points

We now want to enrich the nested quadrature rule $Q_{2\mu+1}^D$ by adding ν generators, that is 2ν points $\pm\lambda_{\mu+1}, \pm\lambda_{\mu+2}, \pm\lambda_{\mu+\nu}$. Observe that the nodal polynomial of the enriched formula can be rewritten as

$$\pi_{2\mu+1+2\nu} = \pi_{2\mu+1}\pi_{2\nu}, \qquad \pi_{2\mu+1} = t\prod_{j=1}^{\mu}(t^2 - \lambda_j^2), \quad \pi_{2\nu} = \prod_{j=1}^{\nu}(t^2 - \lambda_{\mu+j}^2)\,,$$

with $\pi_{2\mu+1}$ odd polynomial and $\pi_{2\nu}$ even polynomial.

Equation (4.45) in Theorem 4.1 provides a means for the computation of the optimal generators, by enforcing $\pi_{2\nu}$ to satisfy the following set of ν equations

$$\int_{-\infty}^{\infty} \pi_{2\mu+1}(t)\pi_{2\nu}(t)t^k w(t)dt = 0, \qquad k = 1, 3, \ldots, 2\nu - 1. \tag{4.47}$$

Note indeed that the remaining ν equations with even k, $k = 0, 2, \ldots, 2\nu - 2$ are identically true, having chosen $\pi_{2\nu}$ as an even polynomial with symmetric roots (recall that $\pi_{2\mu+1}$ is odd and w is even).

Equation (4.47) is a set of non linear equations in λ_j. However, it can be recast into a linear system for the coefficients a_k such that $\pi_{2\nu} = t^{2\nu} + \sum_{j=1}^{\nu} a_j t^{2\nu-2j}$. Finally, a root-finding procedure will compute the new generators $\lambda_{\mu+1}, \lambda_{\mu+2}, \lambda_{\mu+\nu}$.

The next two Lemmas state respectively a condition on the minimum number of points that has to be added and the degree of exactness of the resulting rule.

Lemma 4.4. *Given a quadrature rule $Q^D_{2\mu+1}$, (4.47) is well-posed only if*

$$2\nu > D - (2\mu + 1). \tag{4.48}$$

Proof. Note that $\pi_{2\nu} t^k$ is a polynomial with degree $2\nu + k$. For $Q^D_{2\mu+1}$ Theorem 4.1 holds with $D = (2\mu+1)+m-1$, so that $\int \pi_{2\mu+1}(t)p(t)w(t)dt$ is identically zero for all polynomials with degree up to $m - 1 = D - (2\mu + 1)$. Thus (4.47) is identically true unless $2\nu + k > D - (2\mu + 1)$ for all $k = 0, \dots, 2\nu - 1$. □

Corollary 4.2. *If $Q^D_{2\mu+1}$ is a Gaussian rule,* (4.47) *is well-posed only if $\nu > \mu$.*

Proof. A $(2\mu + 1)$-Gaussian rule has degree of exactness $D = 2N - 1 = 2(2\mu + 1) - 1 = 4\mu + 1$. Thus (4.48) in Lemma 4.4 implies $2\nu > 4\mu + 1 - (2\mu + 1) = 2\mu$. □

Lemma 4.5. *Given a symmetric density function w, the enriched rule $Q^E_{2\mu+1+2\nu}$ obtained adding 2ν symmetric points to $Q^D_{2\mu+1}$ with odd modal polynomial $\pi_{2\mu+1}$ has degree of exactness $E = 4\nu + (2\mu + 1)$.*

Proof. Observe that the symmetry hypotheses imply that also the $2\nu + 1$-th equation of system (4.47) holds true. Therefore Theorem 4.1 holds for the new quadrature rule with $m-1 = 2\nu$, hence $E = N + (m - 1) = N + 2\nu = (2\mu + 1 + 2\nu) + 2\nu$. □

Remark 4.5. *Consider again a Gaussian quadrature rule $Q^D_{2\mu+1}$. Corollary 4.2 states that one has to add at least 2ν points, with $\nu > \mu$, which results in a rule with $D = 4\nu + (2\mu + 1)$ according to Lemma 4.5. A natural question that arises is whether the new formula is actually more accurate than the original Gaussian quadrature, that has degree of exactness $D = 2(2\mu + 1) - 1$ according to Corollary 4.1. In other words, one may ask if the minimum addition $\nu > \mu$ is enough to get $4\nu + (2\mu + 1) > 2(2\mu + 1) - 1$. The latter is equivalent to $4\nu + 2\mu > 4\mu$, hence $\nu > \mu/2$. Thus the minimal addition is indeed enough to actually improve the previous Gaussian quadrature rule, and* a fortiori *enough to improve a non Gaussian quadrature rule, which has a lower degree of exactness.*

We are now in position to build the *KPN* quadrature rule. We first consider the one-point Gauss–Hermite Q^1_1 ($\mu = 0$). Adding 2 points, $\nu = 1$, one obtains the three-points Gauss–Hermite quadrature rule $Q^5_3[1 + 2]$ ($\mu = 1$).

The next step would be to add 2 more generators (4 points). Solving (4.47) one gets to $\pi_{2\nu} = t^4 - 10t^2 - 5$, which however has complex roots and thus cannot be used to exted $Q^5_{3,\mu=1}[1 + 2]$. On the other hand, adding 6 points ($\nu = 3$) leads to a nodal polynomial $\pi_{2\nu}$ with real roots, that can be therefore added to $Q^5_3[1 + 2]$. The resulting rule has 9 points and degree of exactness $D = 4\nu + 2\mu + 1 = 4 \cdot 2 + 2 \cdot 1 + 1 = 15$, $Q^{15}_9[1 + 2 + 6]$. This process can be repeated; the next quadrature rules found are $Q^{29}_{19}[1 + 2 + 6 + 10]$ and $Q^{51}_{35}[1 + 2 + 6 + 10 + 16]$.

Starting from Q^1_1, $Q^5_3[1+2]$, it is possible to derive an alternative chain of quadrature rules that also leads to a quadrature rule with degree 51: Q^1_1, $Q^5_3[1+2]$, $Q^{19}_{11}[1+2+8]$, $Q^{51}_{31}[1+2+8+20]$. Such family however was not used in this work, since the addition of 20 points in a single quadrature level would lead to an excessive cost for the sparse grid procedure.

Chapter 5

Generalized stochastic spectral decomposition for the steady Navier-Stokes equations

This Chapter mainly consists of the paper by L. Tamellini, O.P. Le Maitre, A. Nouy, *Generalized stochastic spectral decomposition for the Steady Navier–Stokes equations*, currently in preparation.

5.1 Introduction

In this Chapter we focus again on the Stochastic Galerkin method (see Chapters 2 and 3), with the aim to present a further technique to reduce its computational costs, the so-called *Proper Generalized Decomposition* (PGD). Such strategy is an alternative to the Optimal Sets approach for the Galerkin method presented in 3, and can be applied to a broad class of problems, including non-elliptic and even non-linear ones.

The computational cost of the Galerkin method is mainly due to the number of probabilistic orthogonal polynomials needed to span the subspace where the solution is sought. While the Optimal Sets approach reduces the cardinality of such basis by tailoring the polynomial subspace to the problem at hand, the *PGD* approach instead looks for a basis in terms of general stochastic polynomials, not necessarily orthogonal. The *PGD* method indeed looks for an approximation of the solution given by $u \approx \sum_{i=1}^{m} u_i \lambda_i$, where u_i are deterministic functions, λ_i are generic polynomials over the probabilistic subspace, and nor u_i neither λ_i are fixed “a-priori”. The *PGD* approach thus can be seen as a reduced basis technique; see e.g. [78] for a literature survey on earlier attempts to introduce reduced approximations in the context of PDEs with stochastic coefficients.

Note that if the solution u was known, the *PGD* expansion could be obtained by computing the Karhunen-Loève expansion. However, since u is of course unknown, what is needed is an algorithm that computes the *PGD* expansion based on the equations solved by the solution rather than the solution itself.

It was first shown in [75, 74] that the modes U_i, λ_i are solution of an eigen-like problem, and therefore the modes may be computed with power-iteration methods. One of the main advantages of such iterative methods is to separate the resolution of the deterministic and stochastic problems, so that one can reuse the existing deterministic solvers with minimal adaptations.

These methods have been successfully applied to scalar problems, both linear [74, 75] and non-linear [78] as well as to time-dependent problems [74]; see also [76] for an earlier attempt on extending *PGD* to non-linear problems. In this Chapter we thus focus on vector, non-linear problems, and in particular on the steady-state Navier–Stokes equations with uncertainty on the viscosity parameter and on the forcing term.

The rest of this Chapter is organized as follows: Section 5.2 will detail the *PGD* principles and available power-like Algorithms. The Navier–Stokes equations with uncertainty will be described in Section 5.3, and numerical validation of the method will be presentaed in Section 5.4. In the context of the Navier–Stokes problem, a non-trivial issue is represented by the computation of the stochastic pressure: Section 5.5 is therefore dedicated to such problem. Finally, 5.6 draws some conclusions.

5.2 Proper Generalized Decomposition (PGD)

5.2.a Stochastic variational problem

Consider the abstract deterministic variational problem given by

Find $u \in \mathcal{V}$ such that

$$a\,(u, v\,;\pi) = b(v\,;\pi), \qquad\qquad \forall v \in \mathcal{V}, \tag{5.1}$$

with $\mathcal{V}$ an appropriate vector space, π the problem parameters, and two forms

$$a(u, v\,;\pi) : (u, v) \in \mathcal{V}^2 \to \mathbb{R} \qquad \text{and} \qquad b(v\,;\pi) : v \in \mathcal{V} \to \mathbb{R}.$$

The forms a and b are parametrized by π and assumed linear with regard to the second and first arguments respectively. The deterministic space $\mathcal{V}$ can be here either infinite or finite dimensional and is equipped with an inner product $(\cdot,\cdot)_{\mathcal{V}}$ with associated norm $\|\cdot\|_{\mathcal{V}}$. Note that if $\mathcal{V}$ has infinite dimension, it will have to be discretized at some point. However, to remain as general as possible, we delay the discussion on discretized spaces $\mathcal{V}$ to the next sections. In any case, we assume that problem (5.1) has a unique solution (depending on π).

We are interested in situations where the parameters π of the problem are uncertain and considered as random. Let $\mathcal{P} = (\Omega, \mathcal{F}, P)$ be a probability space, where Ω is the set of random events, $\mathcal{F}$ the σ-algebra of the events and P a probability measure. For π defined on $\mathcal{P}$, we denote by $\pi(\omega)$, $\omega \in \Omega$, a realization of the problem parameters. The expectation of a generic random quantity h defined on $\mathcal{P}$ is denoted

$$\mathbb{E}\,[h] = \int_\Omega h(\omega)\, dP(\omega).$$

Let $\mathrm{L}^2_P(\Omega)$ be the space of second-order random quantities, equipped with the inner product $(\cdot,\cdot)_P$ and associated norm $\|\cdot\|_{\mathrm{L}^2_P(\Omega)}$,

$$(h, g)_P = \int_\Omega h(\omega) g(\omega)\, dP(\omega) \quad \forall (h, g) \in \mathrm{L}^2_P(\Omega), \quad \|h\|_{\mathrm{L}^2_P(\Omega)} = (h, h)_P^{1/2},$$

so that $h \in \mathrm{L}^2_P(\Omega) \;\Leftrightarrow\; \|h\|_{\mathrm{L}^2_P(\Omega)} < +\infty$. Since the parameters π in equation (5.1) are random, its solution u is also random, defined on $\mathcal{P}$, and satisfies equation (5.1) for a.e. $\omega \in \Omega$, that is

Find $U = U(\omega) : \Omega \to \mathcal{V}$ such that

$$a\,(U(\omega), v\,;\pi(\omega)) = b(v\,;\pi(\omega)\,), \qquad\qquad \forall v \in \mathcal{V},\ \textit{for a. e. } \omega \in \Omega. \tag{5.2}$$

It will be further assumed that the stochastic solution $U \in \mathcal{V} \otimes \mathrm{L}^2_P(\Omega)$, so that the fully weak variational form of the stochastic problem is given by

Stochastic problem.

Find $U \in \mathcal{V} \otimes \mathrm{L}^2_P(\Omega)$ *such that*

$$A\left(U, V\,; \pi\right) = B(V\,; \pi), \qquad \forall\ V \in \mathcal{V} \otimes \mathrm{L}^2_P(\Omega), \tag{5.3}$$

where the forms A and B are given by

$$\begin{aligned} A\left(U, V\,; \pi\right) &= \mathbb{E}\left[a\left(U, V\,; \pi\right)\right] = \int_\Omega a\left(U(\omega), V(\omega)\,; \pi(\omega)\right)\, dP(\omega), \\ B(V\,; \pi) &= \mathbb{E}\left[b(V\,; \pi)\right] = \int_\Omega b(V(\omega)\,; \pi(\omega))\, dP(\omega). \end{aligned}$$

5.2.b Stochastic discretization

For the purpose of numerical simulations, numerical discretizations need to be introduced. These will concern both the deterministic space $\mathcal{V}$, to be discussed in the following sections, and the stochastic space $\mathrm{L}^2_P(\Omega)$, for which we rely on generalized Polynomial Chaos Expansions (gPCE).

To this end, we consider a set of N independent identically distributed random variables, $\mathbf{y} = \{y,\ i = 1, \dots, N\}$, defined on $\mathcal{P}$ with range Γ and known probability density function $\rho(\mathbf{y})$. Any functional $h : \mathbf{y} \in \Gamma \to \mathbb{R}$ is then a real-valued random variable and we have

$$\mathbb{E}\left[h\right] = \int_\Omega h(\mathbf{y}(\omega))\, dP(\omega) = \int_\Gamma h(\mathbf{y})\rho(\mathbf{y})d\mathbf{y}, \quad \text{with } \int_\Gamma \rho(\mathbf{y})d\mathbf{y} = 1.$$

We further assume that the random parameters π are functions of $\mathbf{y}$ (see examples in the results sections), that is

$$\pi(\omega) = \pi(\mathbf{y}(\omega)) \text{ a.s.}.$$

Since the model parameters are the only source of stochasticity in the problem, we have $U(\omega) = U(\mathbf{y}(\omega))$ for the solution of (5.2). Consequently, one can compute the solution in the stochastic space $(\Gamma, \mathcal{B}(\Gamma), \rho(\mathbf{y})d\mathbf{y})$ spanned by $\mathbf{y}$, called the image space, instead of in the abstract space $\mathcal{P}$. To this end, we denote $\mathrm{L}^2_\rho(\Gamma)$ the space of second-order random variables, equipped with the inner product and associated norms

$$\langle \lambda, \beta \rangle = \int_\Gamma \lambda(\mathbf{y})\beta(\mathbf{y})\rho(\mathbf{y})d\mathbf{y} = \mathbb{E}\left[\lambda\beta\right] \quad \forall (\lambda, \beta) \in \mathrm{L}^2_\rho(\Gamma)^2, \qquad \|\lambda\|_{\mathrm{L}^2_\rho(\Gamma)} = \langle \lambda, \lambda \rangle^{1/2}.$$

Next, we introduce a N-variate orthonormal polynomial basis for $\mathrm{L}^2_\rho(\Gamma)$, $\{\mathcal{H}_1, \mathcal{H}_2, \dots\}$, and denote by $\mathbb{P}_M(\Gamma)$ the subspace of $\mathrm{L}^2_\rho(\Gamma)$ spanned by the first M elements of the stochastic basis, that is

$$\mathrm{L}^2_\rho(\Gamma) \supset \mathbb{P}_M(\Gamma) = \operatorname{span}\{\mathcal{H}_1, \dots, \mathcal{H}_M\}.$$

An element $\lambda \in \mathrm{L}^2_\rho(\Gamma)$ can be approximated by $\lambda^M \in \mathbb{P}_M(\Gamma)$ defined by the gPCE expansion

$$\lambda^M(\mathbf{y}) = \sum_{i=1}^M \lambda_i \mathcal{H}_i(\mathbf{y}), \qquad \lim_{M\to\infty} \|\lambda^M - \lambda\|_{\mathrm{L}^2_\rho(\Gamma)} = 0.$$

Each standard measure $\rho(\mathbf{y})d\mathbf{y}$ over Γ leads to a different classical polynomial family [111], the case of y_i standard Gaussian random variables corresponding to (normalized) Hermite polynomials [42]. It is remarked that all developments below immediately extend to other types of stochastic discretizations.

For polynomial stochastic basis, a common truncature strategy is based on the maximal total degree of the basis functions retained in the construction of $\mathbb{P}_M(\Gamma)$ (TD spaces in Chapters 2,3). Denoting w the maximal total degree, the dimension of $\mathbb{P}_M(\Gamma)$ is

$$\dim(\mathbb{P}_M(\Gamma)) = M = \frac{(N+w)!}{N!w!},$$

highlighting its combinatoric increase with both the number of random variables in $\mathbf{y}$ and the expansion degree w. Other possible choices for $\mathbb{P}_M(\Gamma)$ have been investigated in Chapters 2,3.

5.2.c Stochastic Galerkin formulation

Having introduced the discretized stochastic space $\mathbb{P}_M(\Gamma)$, the Stochastic problem (5.3) can be recast as

Discrete Stochastic Problem.

Find $U^M \in \mathcal{V} \otimes \mathbb{P}_M(\Gamma)$ *such that*

$$A\left(U^M, V^M; \pi\right) = B(V^M; \pi), \qquad \forall V^M \in \mathcal{V} \otimes \mathbb{P}_M(\Gamma).$$

One classical way of approximating the solution of such a stochastic variational problem is the stochastic Galerkin projection method. Inserting the gPCE expansion $U^M = \sum_{i=1}^{M} u^i \mathcal{H}_i$ in the previous equations results in a set of M coupled problems for the deterministic Galerkin modes u^i of the solution, namely

$$A\left(\sum_{i=1}^{M} u^i \mathcal{H}_i, v^l \mathcal{H}_l; \pi\right) = B(v^l \mathcal{H}_l; \pi), \qquad \forall v \in \mathcal{V} \text{ and } l = 1, \ldots, M. \tag{5.4}$$

It is seen that the dimension of the Galerkin problem is M times larger than the size of the original deterministic problem. Consequently, its resolution can be very costly, or even prohibitive, whenever N or w needs be large to obtain an accurate approximate U^M of the exact stochastic solution. An additional difficulty appears when the form a is nonlinear in its first argument, making difficult the computation of the stochastic form A. These two limitations call for improvement. First, regarding the dimensionality of the Galerkin problem, one can reduce complexity by relying on more appropriate expansion basis, *e.g.* by means of adaptive strategies and enrichment of polynomial basis (see e.g. Chapter 3, and [20, 21]). However, adaptive approaches are in general complex to implement and often remain computationally intensive, while they do not address the difficulties related to nonlinearities. On the contrary, the PGD approaches discussed in the following aim at tackling the issues of dimensionality and, to some extent, are better suited to the reuse of deterministic code with less concerns on nonlinearities as a result. This latter point will be further discussed in the following.

5.2.d PGD: principles

Let us go back to Formulation 5.3. The PGD seeks for a separated representation of the solution $U \in \mathcal{V} \otimes \mathrm{L}^2_\rho(\Gamma)$ as

$$U(\mathbf{y}) = \sum_i u_i \lambda_i(\mathbf{y}),$$

where the $u_i \in \mathcal{V}$ are called the deterministic components and the $\lambda_i \in \mathrm{L}^2_\rho(\Gamma)$ the stochastic components of the PGD. The m-terms PGD approximation of U, denoted $U^{(m)}$, corresponds to the truncated series

$$U^{(m)} = \sum_{i=1}^{m} u_i \lambda_i \approx U. \tag{5.5}$$

The objective is then to construct the truncated expansion (5.5), without prior knowledge of deterministic and stochastic components, to minimize the approximation error. This has to be contrasted with the Galerkin approach where the stochastic components, the $\mathcal{H}_i$, are selected *a priori*, before the computation.

The simplest PGD algorithms determine the couples $(u_i, \lambda_i) \in \mathcal{V} \times \mathrm{L}^2_\rho(\Gamma)$ one after the others. Specifically, assuming that $U^{(m)}$ has been already determined, let us denote (u, λ) the next couple of components. This couple is solution of the following Galerkin problem:

Couple problem.

Find $(u, \lambda) \in \mathcal{V} \times \mathrm{L}^2_\rho(\Gamma)$ *such that*

$$A\left(U^{(m)} + u\lambda, v\beta\,; \pi\right) = B(v\beta\,; \pi), \qquad \forall (v, \beta) \in \mathcal{V} \times \mathrm{L}^2_\rho(\Gamma).$$

We observe that the solution of this problem is not unique, for if (u, λ) is solution then $(\alpha u, \lambda/\alpha)$ is also solution, $\forall \alpha \neq 0$. Using the sought solution (u, λ), one can derive two auxiliary problems. Specifically, if λ was known, $u \in \mathcal{V}$ would solve

Deterministic Problem.

Find $u \in \mathcal{V}$ *such that*

$$A\left(U^{(m)} + u\lambda, v\lambda\,; \pi\right) = B(v\lambda\,; \pi), \qquad \forall v \in \mathcal{V}. \tag{5.6}$$

We denote hereafter $u = \mathscr{D}(\lambda; U^{(m)})$ the solution of this *deterministic* problem. Similarly, if u was known, $\lambda \in \mathrm{L}^2_\rho(\Gamma)$ would solve

Stochastic Problem.

Find $\lambda \in \mathrm{L}^2_\rho(\Gamma)$ *such that*

$$A\left(U^{(m)} + u\lambda, u\beta\,; \pi\right) = B(u\beta\,; \pi), \qquad \forall \beta \in \mathrm{L}^2_\rho(\Gamma). \tag{5.7}$$

We denote hereafter $\lambda = \mathscr{S}(u; U^{(m)})$ the solution of this *stochastic* problem.

5.2.e PGD: algorithms

Since (u, λ) are both unknown, (5.6) and (5.7) cannot be used for the calculation of u and λ respectively. In addition, one wants the couple (u, λ) to effectively reduce the approximation error.

In the case of linear, symmetric, positive definite form A, it was shown in [75] that the sought couples (u, λ) can be interpreted as the solution of a Rayleigh quotient. This interpretation suggests to reuse techniques for the resolution of eigenvalues problems, like power-iteration techniques, to extract the couples (u, λ), see e.g. [74]. Their application to scalar non linear problems has been thoroughly investigated in [78].

Algorithm 2 Power method

```
 1: $U \leftarrow 0$                                    [*element 0 of $\mathcal{V}$*]
 2: for $l$ in $1, 2, \dots, m$ do
 3:    Initialize $\lambda$                             [*e.g. at random*]
 4:    repeat
 5:       Solve deterministic problem: $u \leftarrow \mathscr{D}(\lambda; U)$
 6:       Normalize $u$: $u \leftarrow u/\|u\|_{\mathcal{V}}$
 7:       Solve stochastic problem: $\lambda \leftarrow \mathscr{S}(u; U)$
 8:    until $(u, \lambda)$ converged
 9:    $U \leftarrow U + u\lambda$
10: end for
```

Power-Iterations

The power method for the computation of (u, λ) is stated in Algorithm 2.

Note that the convergence criteria on the couple (u, λ) yielded by the power-type iterations is understood in a broad sense since it may converge to a subspace whose dimension is greater than one (see [74, 75] for discussion on the convergence of the iterations). In practice, only a limited number of iterations is performed. We also remark that λ and u have equivalent roles in the Algorithm, so that e.g. the normalization step at line 6 could be performed on λ rather then u.

The convergence of the resulting PGD obtained by the Power-Iteration algorithm can be improved by introducing an update of the stochastic components $\{\lambda_1, \dots, \lambda_m\}$ after the determination of the m-th first couples. More specifically, given the deterministic components $u_1, u_2, \dots, u_m$, the update problem consists in the solution of the following set of m coupled equations:

Update problem.

Find $\lambda_1, \dots, \lambda_m$ *such that*

$$A\left(\sum_{i=1}^{m} u_i\lambda_i, u_l\beta\,; \pi\right) = B(u_l\beta\,; \pi), \qquad \forall \beta \in \mathrm{L}^2_\rho(\Gamma),\ l = 1, \dots, m. \tag{5.8}$$

Denoting $\Lambda^{(m)} = [\lambda_1 \dots \lambda_m]$, the update problem is compactly written formally as

$$\Lambda^{(m)} = \mathscr{U}(\mathbf{W}^{(m)}),$$

where $\mathbf{W}^{(m)} = [u_1 \dots u_m]$ is called the reduced deterministic basis (of $\mathcal{V}$). The power-type algorithm with update is stated in Algorithm 3.

Note that it is not necessary to solve the update problem (line 13 of Algorithm 3) at every step l. Moreover, it would be possible to update $\mathbf{W}$ instead of Λ, given the simmetry of the Algorithm with respect to the deterministic and stochastic modes; in this case the normalization step at line 8 should be performed on λ rather than u.

Arnoldi iterations

One disadvantage of Power-iterations-like methods is that they discard all the intermediate solutions within the *repeat-until* loops. The so-called Arnoldi algorithm is a possible solution to overcome such a “waste”: the temporary solutions are used to build a deterministic orthogonal basis $\mathbf{W}^{(m)}$, and then an update problem is solved to compute $\Lambda^{(m)}$. The main advantage of this algorithm is therefore that it requires a lower number of resolutions for the determinstic and stochastic problems. The Arnoldi algorithm is stated in Algorithm 4.

Algorithm 3 Power method with update

```
$U \leftarrow 0$                                              [*element 0 of $\mathcal{V}$*]
$\mathbf{W} \leftarrow \{\}$                                              [*initialization of the reduced basis for $u$*]
$\Lambda \leftarrow \{\}$                                              [*initialization of the reduced basis for $\lambda$*]
for $l$ in $1, 2, \ldots, m$ do
    Initialize $\lambda$                                      [*e.g. at random*]
    repeat
        Solve deterministic problem: $u \leftarrow \mathscr{D}(\lambda; U^{(l)})$
        Normalize $u$: $u \leftarrow u/\|u\|_{\mathcal{V}}$
        Solve stochastic problem: $\lambda \leftarrow \mathscr{S}(u; U^{(l)})$
    until $(u, \lambda)$ converged
    Add $u$ to its reduced basis: $\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l-1)} \cup \{u\}$
    Add $\lambda$ to its reduced basis: $\Lambda^{(l)} \leftarrow \Lambda^{(l-1)} \cup \{\lambda\}$
    Solve update problem: $\Lambda^{(l)} \leftarrow \mathscr{U}(\mathbf{W}^{(l)})$
    $U^{(l)} \leftarrow \sum_{k=1}^{l} u_k \lambda_k$
end for
```

Algorithm 4 Arnoldi method

```
$l \leftarrow 0$                                              [*initialize counter for modes*]
$\mathbf{W} \leftarrow \{\}$                                              [*void container for deterministic modes*]
$\Lambda \leftarrow \{\}$                                              [*void container for stochastic modes*]
$U \leftarrow 0$                                              [*element 0 of $\mathcal{V}$*]
Initialize $\lambda$                                          [*e.g. at random*]
while $l < m$ do
    $l \leftarrow l + 1$
    Solve deterministic problem $u^* \leftarrow \mathscr{D}(\lambda; U)$
    Orthogonalize $u^*$: $u \leftarrow u^* - \sum_{k=1}^{l-1} (u_k, u^*)_{\mathcal{V}}$
    if $\|u\|_{\mathcal{V}} < \varepsilon$ then
        $l \leftarrow l - 1$                                  [*stagnation of Arnoldi detected*]
        Solve update problem: $\Lambda \leftarrow \mathscr{U}(\mathbf{W})$
        $U \leftarrow \sum_{k=1}^{l} u_k \lambda_k$
    else
        Normalize $u$: $u \leftarrow u/\|u\|_{\mathcal{V}}$
        Solve stochastic problem: $\lambda \leftarrow \mathscr{S}(u; U)$
        Add $u$ to its container: $\mathbf{W} \leftarrow \mathbf{W} \cup \{u\}$
        Add $\lambda$ to its container: $\Lambda \leftarrow \Lambda \cup \{\lambda\}$
        if $l = m$ then
            Solve update problem: $\Lambda \leftarrow \mathscr{U}(\mathbf{W})$
            $U \leftarrow \sum_{k=1}^{l} u_k \lambda_k$
        end if
    end if
end while
```

Whenever the generation of deterministic modes stagnates onto invariant subspaces (detected using the small positive parameter ε at line 10) an update step is performed. Note also that the update problems at lines 12 and 20 concern the whole stochastic components $\Lambda^{(l)}$ generated so far, but one could as well perform a partial update considering only the Arnoldi subspace generated after the last detected stagnation.

5.2.f Practical considerations

Obviously, also the algorithms above need a stochastic discretization. Again, we shall rely on gPCE expansions for the stochastic components and approximate the stochastic modes λ_i in the finite dimensional $\mathbb{P}_M(\Gamma)$ by $\sum_{k=0}^{M} \lambda_i^k \mathcal{H}_k$. Further, with this stochastic discretization, the stochastic (5.7) and update (5.8) problems translate to the Galerkin problems

$$A\left(U^{(m)} + u \sum_{k=1}^{M} \lambda_i^k \mathcal{H}_k, u\mathcal{H}_l\,;\pi\right) = B(u\mathcal{H}_l\,;\pi), \qquad l = 1,\dots,M, \tag{5.9}$$

and

$$A\left(\sum_{i=1}^{m} u_i \left(\sum_{k=1}^{M} \lambda_i^k \mathcal{H}_k\right), u_l\mathcal{H}_j\,;\pi\right) = B(u_l\mathcal{H}_j\,;\pi), \qquad l = 1,\dots,m \text{ and } j = 1,\dots,M. \tag{5.10}$$

For a given stochastic approximation space $\mathbb{P}_M(\Gamma)$, one can expect that the PGD solution $U^{(m)}$ to converge quickly to the Galerkin solution $U^M \in \mathcal{V} \otimes \mathbb{P}_M(\Gamma)$, with $m \ll M$ modes. This expectation comes from the fact that the PGD constructs the more relevant stochastic components λ_i for the expansion, contrary to the Galerkin case where one chooses *a priori* the stochastic components (as the elements of the gPCE basis) and then seek for the solution in $\mathbb{P}_M(\Gamma)$.

Another point to be underlined in view of the above algorithms is that in each of them the computationally intensive steps are the resolution of the deterministic and stochastic problems, plus the update problems (optional in the Power-Iteration algorithm). As seen in (5.6) and (5.9) the size of the deterministic and stochastic problems are constant and equal to the dimension of the discretized spaces $\mathcal{V}$ and $\mathbb{P}_M(\Gamma)$ respectively; this is in general much lower than the size of the Galerkin problem which is the product of the two, with a significant complexity reduction as a result (provided that the number of systems to be solved is small enough). Concerning the update problem, we observe that its dimension is $m \times \dim(\mathbb{P}_M(\Gamma))$ so that if m is less than the dimension of the discretized space $\mathcal{V}$ the update problem is again much smaller in size than the Galerkin problem.

In addition, it will be shown in the following sections that for the Navier-Stokes equations the actual deterministic problems to be solved have structures very similar to the original Navier-Stokes equations, facilitating the re-use of existing deterministic codes, while implementing a Galerkin solver would require a greater implementation effort.

We also remark that instead of updating the stochastic components of the PGD solution, one could instead derive an update problem for the deterministic components $\{u_i, i = 1,\dots,m\}$, which would in fact have the structure of the Galerkin problem in (5.4) but for the approximation in the stochastic space spanned by the $\{\lambda_i\}$ instead of the $\{\mathcal{H}_i\}$. This alternative should be considered for problems where the dimension M of the stochastic space exceeds that of the discretized space $\mathcal{V}$.

5.3 Navier-Stokes equations with uncertain inputs

We consider the bidimensional, steady, incompressible (constant density) Navier-Stokes equations on a bounded, simply connected domain $D \subset \mathbb{R}^2$ with boundary ∂D. The dimensionless Navier-

Stokes equations are

$$\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{u} = -\boldsymbol{\nabla} p + \boldsymbol{\nabla} \cdot \overline{\sigma}(\boldsymbol{u}) + \boldsymbol{f}, \tag{5.11a}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0, \tag{5.11b}$$

where $\boldsymbol{u} : D \to \mathbb{R}^2$ is the velocity field, $p : D \to \mathbb{R}$ is the pressure field, $\boldsymbol{f} : D \to \mathbb{R}^2$ is the external force field and $\overline{\sigma}$ the viscous stress tensor. For a Newtonian fluid, $\overline{\sigma}$ in (5.11a) has for expression

$$\overline{\sigma}(\boldsymbol{u}) = \frac{\nu}{2} \left(\boldsymbol{\nabla} \boldsymbol{u} + \boldsymbol{\nabla} \boldsymbol{u}^T \right),$$

where $\nu > 0$ is the viscosity parameter (inverse of a Reynolds number), measuring relative influence of the inertial (nonlinear) and viscous (linear) contributions. Accounting for the mass conservation equation (5.11b), the Navier-Stokes equations reduce to

$$\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{u} = -\boldsymbol{\nabla} p + \nu \boldsymbol{\nabla}^2 \boldsymbol{u} + \boldsymbol{f}, \tag{5.12a}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0. \tag{5.12b}$$

These equations have to be complemented with boundary conditions; for simplicity, we shall restrict ourselves to the case of homogeneous Dirichlet velocity boundary conditions on ∂D,

$$\boldsymbol{u}(\mathbf{x}) = 0, \quad \mathbf{x} \in \partial D. \tag{5.13}$$

Next, we classically denote by $\mathrm{L}^2(D)$ the space of functions that are square integrable over D, and are equipped with the inner product and norms

$$(p, q) = \int_D p\, q \, d\mathbf{x}, \qquad \|q\|_{\mathrm{L}^2(D)} = (q, q)^{1/2},$$

and define the constrained space

$$\mathrm{L}_0^2(D) = \left\{ q \in \mathrm{L}^2(D) : \int_D q \, d\mathbf{x} = 0 \right\}.$$

Then, let $\mathbf{H}^1(D)$ be the Sobolev space of vector valued functions with all components and their first derivatives being square integrable over D, and $\mathbf{H}_0^1(D)$ the constrained space of such vector functions vanishing on ∂D,

$$\mathbf{H}_0^1(D) = \left\{ \boldsymbol{v} \in \mathbf{H}^1(D), \ \boldsymbol{v} = 0 \text{ on } \partial D \right\}.$$

With this notation the Navier-Stokes system (5.12) with boundary conditions (5.13) is then equivalent to the variational problem

Navier–Stokes equations.

Find $(\boldsymbol{u}, p) \in \mathbf{H}_0^1(D) \times \mathrm{L}_0^2(D)$ *such that*

$$\begin{aligned} &c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}) + \nu\, v(\boldsymbol{u}, \boldsymbol{v}) + d(p, \boldsymbol{v}) = b(\boldsymbol{v}), && \forall \boldsymbol{v} \in \mathbf{H}_0^1(D) \\ &d(q, \boldsymbol{u}) = 0, && \forall q \in \mathrm{L}_0^2(D), \end{aligned} \tag{5.14}$$

with the forms defined by

$$c(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{v}) = \int_D (\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{w}) \cdot \boldsymbol{v} \, d\mathbf{x}, \qquad v(\boldsymbol{u}, \boldsymbol{v}) = \int_D \boldsymbol{\nabla} \boldsymbol{u} : \boldsymbol{\nabla} \boldsymbol{v} \, d\mathbf{x},$$

$$d(p, \boldsymbol{v}) = -\int_D p \boldsymbol{\nabla} \cdot \boldsymbol{v} \, d\mathbf{x}, \qquad b(\boldsymbol{v}) = \int_D \boldsymbol{f} \cdot \boldsymbol{v} \, d\mathbf{x}.$$

The pressure unknown can also be formally suppressed in this weak formulation, by introducing the subspace of divergence-free functions of $\mathbf{H}_0^1(D)$, denoted hereafter $\mathbf{H}_{0,div}^1(D)$,

$$\mathbf{H}_{0,div}^1(D) = \left\{\boldsymbol{v} \in \mathbf{H}_0^1(D),\ \boldsymbol{\nabla} \cdot \boldsymbol{v} = 0 \text{ in } D\right\}.$$

Seeking $\boldsymbol{u} \in \mathbf{H}_{0,div}^1(D)$, the weak form simplifies to

Divergence-free Navier–Stokes equations.

Find $\boldsymbol{u} \in \mathbf{H}_{0,div}^1(D)$ *such that*

$$c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}) + \nu\, v(\boldsymbol{u}, \boldsymbol{v}) = b(\boldsymbol{v}), \qquad \forall \boldsymbol{v} \in \mathbf{H}_{0,div}^1(D). \tag{5.15}$$

Finally, we introduce the uncertain parameters. In this paper, we are concerned by situations where the external forcing $\boldsymbol{f}$ and viscous parameter ν are uncertain and, consistently with the previous section, are seen as functions of a set of N normalized Gaussian variables with zero mean and unit variance, $\nu = \nu(\mathbf{y})$ and $\boldsymbol{f} = \boldsymbol{f}(\mathbf{x}, \mathbf{y})$. As a consequence, the divergence-free Navier–Stokes equation (5.15) has now a stochastic solution $\boldsymbol{u}(\mathbf{y})$. We can therefore state the following formulation:

Find $\boldsymbol{u} = \boldsymbol{u}(\mathbf{y}) : \Gamma \to \mathbf{H}_{0,div}^1(D)$ *such that*

$$c(\boldsymbol{u}(\mathbf{y}), \boldsymbol{u}(\mathbf{y}), \mathbf{V}) + \nu(\mathbf{y})\, v(\boldsymbol{u}(\mathbf{y}), \mathbf{V}) = b(\mathbf{V}\,; \boldsymbol{f}(\mathbf{y})), \qquad \forall\, \mathbf{V} \in \mathbf{H}_{0,div}^1(D),\ \textit{for a.e. } \mathbf{y} \in \Gamma,$$

whose fully weak counterpart can be written immediately as

Stochastic Navier–Stokes problem.

Find $\boldsymbol{u} \in \mathbf{H}_{0,div}^1(D) \otimes \mathrm{L}_\rho^2(\Gamma)$ *such that*

$$C(\boldsymbol{u}, \boldsymbol{u}, \mathbf{V}) + V_\nu(\boldsymbol{u}, \mathbf{V}) = B(\mathbf{V}), \qquad \forall \mathbf{V} \in \mathbf{H}_{0,div}^1(D) \otimes \mathrm{L}_\rho^2(\Gamma). \tag{5.16}$$

The forms C, V_ν and B are given by

$$C(\boldsymbol{u}, \boldsymbol{w}, \mathbf{V}) = \mathbb{E}\left[c(\boldsymbol{u}, \boldsymbol{w}, \mathbf{V})\right], \qquad V_\nu(\boldsymbol{u}, \mathbf{V}) = \mathbb{E}\left[\nu\, v(\boldsymbol{u}, \mathbf{V})\right], \qquad B(\mathbf{V}) = \mathbb{E}\left[b(\mathbf{V}\,; \boldsymbol{f})\right].$$

The previous formulation is ready to be discretized with the Stochastic Galerkin method, introducing the discretized stochastic space $\mathbb{P}_M(\Gamma)$ as in section 5.2.c. In practice, the divergence-free costraint is treated by adding a stochastic pressure field $P(\mathbf{y})$, see e.g. [60]. We will however base the following discussion on PGD on this formulation since we are looking for a PGD decomposition of $\boldsymbol{u}$. We will return back to the issue of pressure later on.

5.3.a PGD Formulation

We now detail the deterministic, stochastic and update problems associated to the iterations of the PGD algorithms.

Deterministic problem

We assume that the m-terms reduced approximation $\boldsymbol{u}^{(m)} = \sum_{i=1}^m \boldsymbol{u}_i \lambda_i$ has been computed; the deterministic PGD problem $\mathscr{D}(\lambda; U^{(m)})$ for the next deterministic mode $\boldsymbol{u}$ given the stochastic mode λ is

Find $\boldsymbol{u} \in \mathbf{H}^1_{0,div}(D)$ *such that*

$$\begin{aligned} C(\lambda\boldsymbol{u},\lambda\boldsymbol{u},\lambda\boldsymbol{v}) &+ C(\lambda\boldsymbol{u},\boldsymbol{u}^{(m)},\lambda\boldsymbol{v}) + C(\boldsymbol{u}^{(m)},\lambda\boldsymbol{u},\lambda\boldsymbol{v}) + V_\nu(\lambda\boldsymbol{u},\lambda\boldsymbol{v}) \\ &= B(\lambda\boldsymbol{v}) - V_\nu(\boldsymbol{u}^{(m)},\lambda\boldsymbol{v}) - C(\boldsymbol{u}^{(m)},\boldsymbol{u}^{(m)},\lambda\boldsymbol{v}), \qquad \forall\boldsymbol{v} \in \mathbf{H}^1_{0,div}(D). \end{aligned}$$

For convenience and to stress the deterministic character of this problem we rewrite it as

Find $\boldsymbol{u} \in \mathbf{H}^1_{0,div}(D)$ *such that*

$$\begin{aligned} c(\boldsymbol{u},\boldsymbol{u},\boldsymbol{v}) + c\left(\boldsymbol{u},\boldsymbol{v}_c^{(m)}(\lambda),\boldsymbol{v}\right) + c\left(\boldsymbol{v}_c^{(m)}(\lambda),\boldsymbol{u},\boldsymbol{v}\right) + \widetilde{\nu}\, v(\boldsymbol{u},\boldsymbol{v};\lambda) \\ = \widetilde{b}(\boldsymbol{v};\boldsymbol{u}^{(m)},\lambda), \forall\boldsymbol{v} \in \mathbf{H}^1_{0,div}(D). \end{aligned} \tag{5.17}$$

In the previous equation we have denoted

$$\boldsymbol{v}_c^{(m)}(\lambda) = \sum_{i=1}^{m} \frac{\mathbb{E}\left[\lambda^2\lambda_i\right]}{\mathbb{E}\left[\lambda^3\right]}\boldsymbol{u}_i\,, \qquad \widetilde{\nu} = \frac{\mathbb{E}\left[\nu\lambda^2\right]}{\mathbb{E}\left[\lambda^3\right]}$$

$$\widetilde{b}(\boldsymbol{v};\boldsymbol{u}^{(m)},\lambda) = \frac{\mathbb{E}\left[\lambda\, b(\boldsymbol{v}\,;\boldsymbol{f})\right]}{\mathbb{E}\left[\lambda^3\right]} - \sum_{i=1}^{m}\frac{\mathbb{E}\left[\lambda\nu\lambda_i\right]}{\mathbb{E}\left[\lambda^3\right]}v(\boldsymbol{u}_i,\boldsymbol{v}) - \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\mathbb{E}\left[\lambda\lambda_i\lambda_j\right]}{\mathbb{E}\left[\lambda^3\right]}c(\boldsymbol{u}_i,\boldsymbol{u}_j,\boldsymbol{v})\,.$$

It is therefore seen that the structure of the deterministic PGD problem is essentially the same as the weak formulation of the deterministic incompressible Navier-Stokes equations, with a few remarkable differences. In particular: i) we have two new convective terms, whose convective velocity is given by $\boldsymbol{v}_c^{(m)}$; ii) the viscosity parameter is different, since its value is now $\widetilde{\nu} = \mathbb{E}\left[\nu\lambda^2\right]/\mathbb{E}\left[\lambda^3\right]$; iii) the forcing term contains all the information about the previous modes which have been already computed.

As a result, the resolution of this problem can re-use existing deterministic flow solvers with minimal adaptation for the computation of the right-hand-side and the additional convection term. In addition, the enforcement of divergence free character of $\boldsymbol{u}$ can be achieved by introducing the deterministic Lagrange multiplier $p \in \mathrm{L}^2_0(D)$.

Stochastic problem

Let us assume again that the m-terms reduced approximation has been computed; the stochastic PGD problem $\mathscr{S}(\boldsymbol{u};U^{(m)})$ for the next stochastic mode λ given the deterministic mode $\boldsymbol{u}$ is

Find $\lambda \in \mathbb{P}_M(\Gamma)$ *such that*

$$\begin{aligned} C(\lambda\boldsymbol{u},\lambda\boldsymbol{u},\beta\boldsymbol{u}) &+ C(\boldsymbol{u}^{(m)},\lambda\boldsymbol{u},\beta\boldsymbol{u}) + C(\lambda\boldsymbol{u},\boldsymbol{u}^{(m)},\beta\boldsymbol{u}) + V_\nu(\lambda\boldsymbol{u},\beta\boldsymbol{u}) \\ &= B(\beta\boldsymbol{u}) - C(\boldsymbol{u}^{(m)},\boldsymbol{u}^{(m)},\beta\boldsymbol{u}) - V_\nu(\boldsymbol{u}^{(m)},\beta\boldsymbol{u}) \qquad \forall\beta \in \mathbb{P}_M(\Gamma). \end{aligned}$$

This is a quadratic equation for λ in weak form. We can highlight this by recasting the previous formulation as

Find $\lambda \in \mathbb{P}_M(\Gamma)$ *such that*

$$\begin{aligned} &\mathbb{E}\left[\lambda^2\beta\right]c(\boldsymbol{u},\boldsymbol{u},\boldsymbol{u}) + \sum_{i=1}^{m}\mathbb{E}\left[\lambda\lambda_i\beta\right]\left(\,c(\boldsymbol{u}_i,\boldsymbol{u},\boldsymbol{u}) + c(\boldsymbol{u}_i,\boldsymbol{u},\boldsymbol{u})\,\right) + \mathbb{E}\left[\nu\lambda\beta\right]v(\boldsymbol{u},\boldsymbol{u}) = \\ &\mathbb{E}\left[\beta\, b(\boldsymbol{u}\,;\boldsymbol{f})\right] - \sum_{i,j=1}^{m}\mathbb{E}\left[\lambda_i\lambda_j\beta\right]c(\boldsymbol{u}_i,\boldsymbol{u}_j,\boldsymbol{u}) - \sum_{i=1}^{m}\mathbb{E}\left[\nu\lambda_i\beta\right]v(\boldsymbol{u}_i,\boldsymbol{u}) \quad \forall\beta \in \mathbb{P}_M(\Gamma). \end{aligned} \tag{5.18}$$

To actually compute the gPCE expansion of λ in $\mathbb{P}_M(\Gamma)$, $\lambda = \sum_{k=0}^{M} \widehat{\lambda}_k \mathcal{H}_k$, one has next to choose $\beta = \mathcal{H}_l$ in (5.18) and solve the following set of M quadratic equations in the coefficients $\widehat{\lambda}_k$:

$$
\begin{aligned}
c(\boldsymbol{u},\boldsymbol{u},\boldsymbol{u}) \sum_{k,k'=1}^{M} \widehat{\lambda}_k \widehat{\lambda}_{k'}\, \mathbb{E}\left[\mathcal{H}_k \mathcal{H}_{k'} \mathcal{H}_l\right] + \sum_{i=1}^{m} \left(c(\boldsymbol{u}_i,\boldsymbol{u},\boldsymbol{u}) + c(\boldsymbol{u}_i,\boldsymbol{u},\boldsymbol{u}) \right) \sum_{k,k'=1}^{M} \widehat{\lambda}_k \widehat{\lambda}_{i,\,k'} \mathbb{E}\left[\mathcal{H}_k \mathcal{H}_{k'} \mathcal{H}_l\right] \\
+ v(\boldsymbol{u},\boldsymbol{u}) \sum_{k,k'=1}^{M} \widehat{\lambda}_k\, \widehat{\nu}_{k'}\, \mathbb{E}\left[\mathcal{H}_k \mathcal{H}_{k'} \mathcal{H}_l\right] \; = \\
\sum_{k'=1}^{M} b\,(\widehat{\boldsymbol{f}}_{k'},\boldsymbol{u})\, \mathbb{E}\left[\mathcal{H}_{k'} \mathcal{H}_l\right] - \sum_{i,j=1}^{m} c(\boldsymbol{u}_i,\boldsymbol{u}_j,\boldsymbol{u}) \sum_{k,k'=1}^{M} \widehat{\lambda}_{i,\,k}\, \widehat{\lambda_{j}}_{,\,k'} \mathbb{E}\left[\mathcal{H}_k \mathcal{H}_{k'} \mathcal{H}_l\right] \\
- \sum_{i=1}^{m} v(\boldsymbol{u}_i,\boldsymbol{u}) \sum_{k,k'=1}^{M} \widehat{\nu}_k \widehat{\lambda}_{i,\,k'} \mathbb{E}\left[\mathcal{H}_k \mathcal{H}_{k'} \mathcal{H}_l\right] \qquad \forall\, l = 1, \ldots M\,,
\end{aligned}
$$

where we have supposed that $\boldsymbol{f}$ admits a gPCE expansion, $\boldsymbol{f}(\mathbf{x},\mathbf{y}) = \sum_{k'=1}^{M} \boldsymbol{f}_{k'}(\mathbf{x})\, \mathcal{H}_{k'}(\mathbf{y})$.

Update Problem

Given the m-terms PGD decomposition of $\boldsymbol{u}$, the update problem recomputes all the m modes λ_i, and consists therefore of m quadratic equations for λ_i, all mutually coupled, but whose structure is close to the stochastic problem (5.18):

Find $\lambda_i \in \mathbb{P}_M(\Gamma), i = 1, \ldots, m$ *such that*

$$
C\left(\sum_{i=1}^{m} \boldsymbol{u}_i \lambda_i, \sum_{i=1}^{m} \boldsymbol{u}_i \lambda_i, \beta \boldsymbol{u}\right) + V\left(\sum_{i=1}^{m} \boldsymbol{u}_i \lambda_i, \beta \boldsymbol{u}\right) = B(\beta \boldsymbol{u}_j),
$$
$$
\forall \beta \in \mathbb{P}_M(\Gamma),\ \forall j = 1, \ldots, m.
$$

As in the Stochastic Problem we take $\beta = \psi_k$, ending up with a quadratic system of equations for $\lambda_1, \ldots, \lambda_m$, whose dimension is therefore $m \times M$.

5.4 Numerical results

In this Section we consider two test cases of increasing complexity and computational cost: in the first one the viscous parameter ν is the only uncertain parameter, while in the second one we consider both the viscous parameter and the forcing term as uncertainty sources. In both cases the aim of the test will be to compare the PGD solution against the Galerkin solution, to assess the effectiveness of the method. All the PGD solutions will be computed with the Arnoldi method described in Section 5.2.e. In particular, the parameter ε to discard the last mode computed and enter the update procedure (line 10) is set to 10^{-2}. Moreover, the actual implementation slightly differs from the one presented in section 5.2.e because of the following details:

- the update procedure is performed only if $\|u\| < \varepsilon$ *and* we have added at least 2 modes from the previous update;
- an early stop (before all the m modes have been computed) is enforced as soon as one of the following conditions holds:
  - the norm of the residual is less than $\texttt{5.e} - \texttt{7}$ (see Section 5.5 for residual computation);
  - the norm of the last λ computed is less than $\texttt{1.e} - \texttt{9}$;

As for the spatial discretization, we will consider a classical Spectral Element Method discretization, see e.g. [17]. In particular, we will use a grid of $\mathtt{Nu} \times \mathtt{Nu}$ Gauss–Lobatto points for the approximation of the components of the velocity, while the pressure is approximated over a $\mathtt{Nu} - 2 \times \mathtt{Nu} - 2$ grid. The non linearity in the Navier–Stokes equation is solved with a preconditioned Quasi-Newton method, and at each step the linear system is solved with a GMRES solver. Once more we remark that the efficiency of the PGD method in determining the reduced approximation of $\mathbf{U}$ does not depend on the Navier–Stokes solver considered, and any technique may be used.

5.4.a Test 1: Random viscosity parameter

In the first test we consider a random viscosity ν given by

$$\nu(\omega) = \overline{\nu} + \nu'(\omega),$$

where $\overline{\nu} > 0$ (ensure the almost sure positivity of ν), and $\nu'(\omega)$ has a Log-normal distribution with median value $\overline{\nu}' > 0$ and coefficient of variation $C_{\nu'} \geq 1$; we further set $\overline{\nu} = 1/100$. For these settings, the random viscosity can be expressed as

$$\nu(\omega) = \overline{\nu} + \overline{\nu}' \exp\left(\sigma y(\omega)\right), \quad \sigma = \frac{\log C_{\nu'}}{2.85}, \tag{5.19}$$

where $y \sim N(0,1)$, ensuring that $\nu' \in [\overline{\nu}'/C_{\nu'}, \overline{\nu}' C_{\nu'}]$ with a probability ≈ 0.995 .

Regarding the deterministic force field, it is well-known that force fields deriving from the gradient of a potential induce no flow for homogeneous boundary conditions. Therefore we consider the deterministic function $\psi(\mathbf{x})$ and define $\boldsymbol{f}$ as

$$\boldsymbol{f} = \boldsymbol{\nabla} \wedge (0,\, 0,\, \psi)^T, \tag{5.20}$$

so that $\boldsymbol{\nabla} \wedge \boldsymbol{f} = (0,\, 0,\, -\nabla^2\psi)^T$. For simplicity we restrict ourselves to forcing terms having constant rotational,

$$\boldsymbol{\nabla} \wedge \boldsymbol{f} = (0,\, 0,\, \Phi)^T, \tag{5.21}$$

and a zero normal component on $\partial\Omega$. This leads to the definition of ψ by

$$\begin{cases} \nabla^2\psi = -\Phi & \text{in } \Omega \\ \psi = 0 & \text{on } \partial\Omega\,. \end{cases} \tag{5.22}$$

It is useful to further define the operator $\mathcal{L} : \mathbf{H}^{-1}(D) \to \mathbf{H}^1_0(D)$ that maps the forcing term Φ in (5.22) to the corresponding solution, that is

$$\mathcal{L}[\Phi] = \psi\,. \tag{5.23}$$

The constant Φ fixes the magnitude of the forcing and is hereafter set to $\Phi = 100\,\overline{\nu}'$ to ensure that $\|\mathbf{U}\|_\Omega \approx 1$. The typical structure of the forcing field $\boldsymbol{f}$ is shown in Figure 5.1. The spatial discretization considered is $\mathtt{Nu} = 51$.

Galerkin solution

We start by setting $\overline{\nu}' = 1/100$, $C_{\nu'} = 1.5$ and $\overline{\nu} = 0.01\overline{\nu}'$ and consider the classical Galerkin Stochastic Projection method for the approximation of $\mathbf{U}$. Guided by the expression of the viscosity in (5.19), we rely on a gPCE expansion of the solution using a single normalized Gaussian random

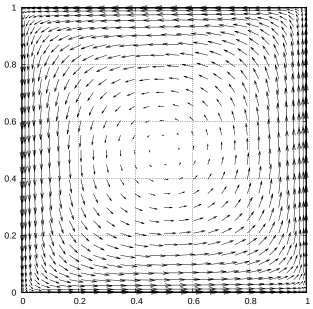


Figure 5.1: Typical structure of the deterministic external forcing used in Section 5.4.

variable y and corresponding Hermite gPCE basis. The Galerkin approximation is therefore sought as

$$\mathbf{U}^G(y) = \sum_{k=0}^{w} \boldsymbol{u}_k^G \mathcal{H}_k(y), \tag{5.24}$$

with w denoting the expansion order and $\mathcal{H}_k$ denoting the k-th degree Hermite polynomial in y. For this random viscosity distribution, a well converged solution is obtained for $w = 10$, as shown in the following discussion.

The Galerkin solution for $w = 10$ is depicted in Figure 5.2, showing the expected velocity field (that is the first mode of the Galerkin solution $\boldsymbol{u}_0^G$, see Figure 5.2(a)), and the expectation and the standard deviation of the rotational of $\mathbf{U}^G$, see Figures 5.2(b) and 5.2(c).

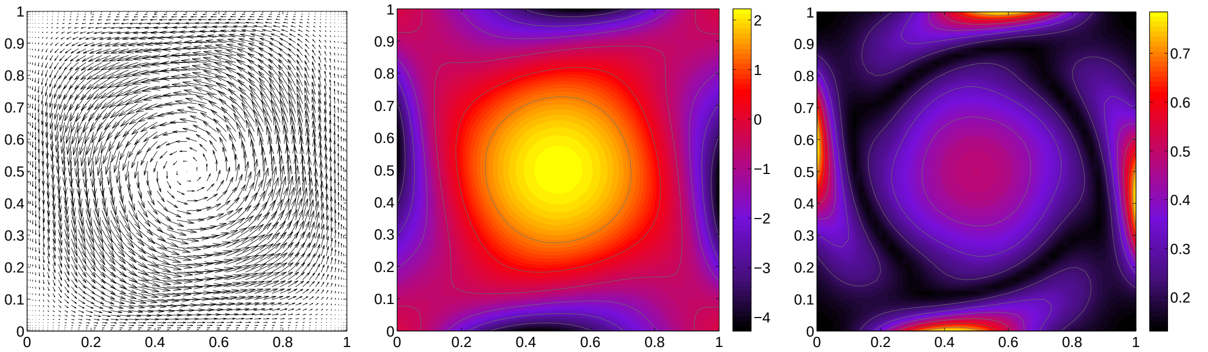
Plots in Figure 5.2 highlight the nonlinear character of the problem. Indeed, for the present situation where the forcing term is deterministic and the viscosity parameter does not depend on $\mathbf{x}$, if the problem was linear the solution would be expressed as a product of a deterministic function times a stochastic factor, $\mathbf{U}(y) = \alpha(y)\boldsymbol{u}^*$, and as a consequence mean and standard deviation of $\mathbf{U}$ would show the same spatial structure, up to a multiplicative factor. This is not the case here, where the expectation and standard deviation field of the velocity rotational exhibit a clearly different spatial pattern. In fact, the random viscosity has the strongest impact on the vorticity field along the boundary of the domain, where the uncertainty level reaches roughly 70%. Another stringent feature of the standard deviation of the vorticity field is the presence of detached structures along the boundary, that are created by the convective effects.

To better appreciate the complexity of the random flow field, as well as the converged character of the Galerkin solution for $w = 10$, the Karhunen-Loève (POD) decomposition of $\mathbf{U}^G(y)$ is computed. Since the Galerkin solution is computed in a subspace $\mathbb{P}_M(\Gamma)$, whose dimension is $w + 1 = 11$, its KL expansion is finite and writes as

$$\mathbf{U}^G(\mathbf{y}) = \sum_{k=0}^{w} \boldsymbol{u}_k^G \mathcal{H}_k(y) = \sum_{l=1}^{w+1} \boldsymbol{u}_l^{G,KL} \sqrt{\kappa_l^G}\eta_l(y), \quad \kappa_1^G \geq \kappa_2^G \geq \cdots \geq \kappa_{w+1}^G \geq 0, \tag{5.25}$$

where $\{\boldsymbol{u}_l^{G,KL}\}$ is an orthonormal set and $\mathbb{E}[\eta_l\eta_{l'}] = \delta_{ll'}$. Figure 5.3 shows the rotational of the KL modes $\boldsymbol{u}_l^{G,KL}$: the plots show the increasing complexity, with the mode index, of the spatial structure of the rotational of the KL modes. They also highlight the impact of the nonlinear convective term which induces a bending of these structures, due to the advection effects, which however possess the symmetries of the present problem.

Figure 5.4 shows the normalized spectrum, that is $S_l = \sqrt{\kappa_l^G / \sum \kappa_n^G}$ for $l = 0, \ldots, w$. It exhibits a fast decay, the 6-th normalized mode being 10^{-6} times the first one, with essentially a uniform



(a) mean velocity field.

(b) mean vorticity field.

(c) standard deviation of the vorticity field.

Figure 5.2: Galerkin solution for Test 1.

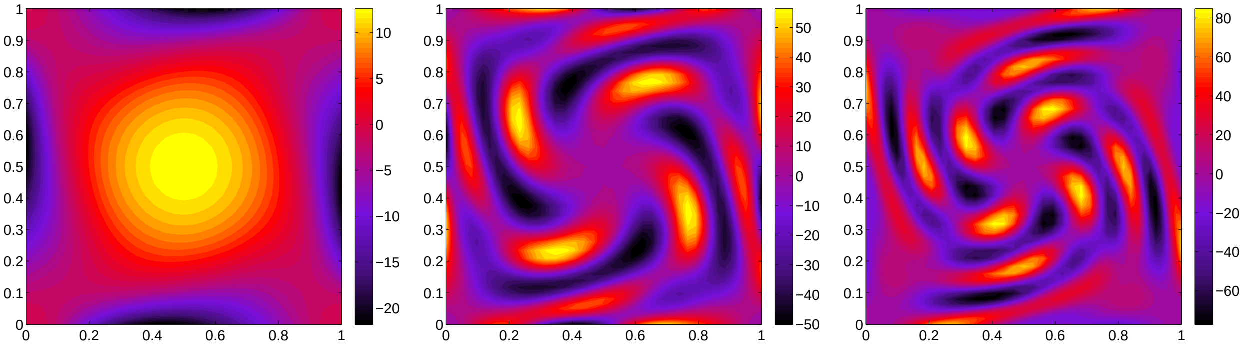


Figure 5.3: Rotational of the KL modes $0, 3, 6$ of the Galerkin solution of Test 1.

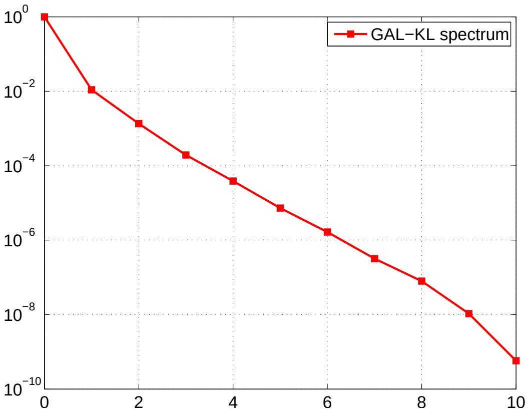


Figure 5.4: KL spectrum of Galerkin solution of Test 1.

asymptotic decay rate except for the very last KL modes which are affected by the truncation of the stochastic basis. The fast decay of the spectrum allows to conclude that the truncation order $w = 10$ is large enough for the present stochastic problem.

PGD solution

We next compute the PGD solution in the same stochastic subspace $\mathbb{P}_M(\Gamma)$ as before, using the Arnoldi algorithm with $\varepsilon = 0.01$ and fixing the maximum number of PGD modes to $m = 15$. Again, we also compute the Karhunen-Loève expansion of the PGD solution, as done for the Galerkin solution.

Figure 5.5 shows the expected velocity field ($\mathbb{E}[\mathbf{U}^{(m)}]$), and the expectation and standard deviation fields of the rotational of $\mathbf{U}^{(m)}$. The plots should be compared with those of the Galerkin solution shown in Figure 5.2, and the agreement is excellent. The same conclusion arises when looking at the rotational of the modes of the KL expansion of the PGD solution, which are shown in Figure 5.6, and have to be compared with Figure 5.3. Figure 5.7 shows the matching between the spectra of the the two KL decompositions (bottom right plot), again showing good agreement between the solutions. Figure 5.8 shows some of the PGD modes of the solution, namely modes $m = 0, 3, 6$, and compares their rotationals with the rotationals of the corresponding modes $m = 0, 3, 6$ of the KL expansion of the PGD solution: the spatial structures are similar, a further confirmation of the accuracy of the PGD representation.

Finally, we investigate the case where the viscosity parameter depends on more than one random variable. To do this, we modify the definiton of ν from equation (5.19) to

$$\nu(\omega) = \overline{\nu} + \exp\left(\frac{\sigma}{\sqrt{N_\nu}} \sum_{i=1}^{N_\nu} y_i(\omega)\right),$$

with y_i independent and identically distributed Gaussian random variables. This is clearly an overparametrization of the problem, since indeed $y_T(\omega) = \sigma/\sqrt{N_\nu}\sum_{i=1}^{N_\nu} y_i(\omega)$ is in turn a Gaussian random variable with zero mean and variance equal to σ^2, therefore ν truly has a unique stochastic dimension, such that the Navier–Stokes solution has the same stochastic dimensionality $\forall N_\nu \geq 1$. It is found that the PGD solution is insensitive to this overparametrization and the deterministic modes it computes are essentially the same, thus proving to be able to capture the key features of the stochastic solution. This clearly appears in Figure 5.9, where we consider the PGD solutions for problems with $N_\nu = 1, 2, 3$: here we compare the norms of the PGD stochastic modes λ_i (Figure 5.9(b)), and the norms of the KL modes of the PGD solutions (Figure 5.9(a)).

5.4.b Test 2: Random forcing term

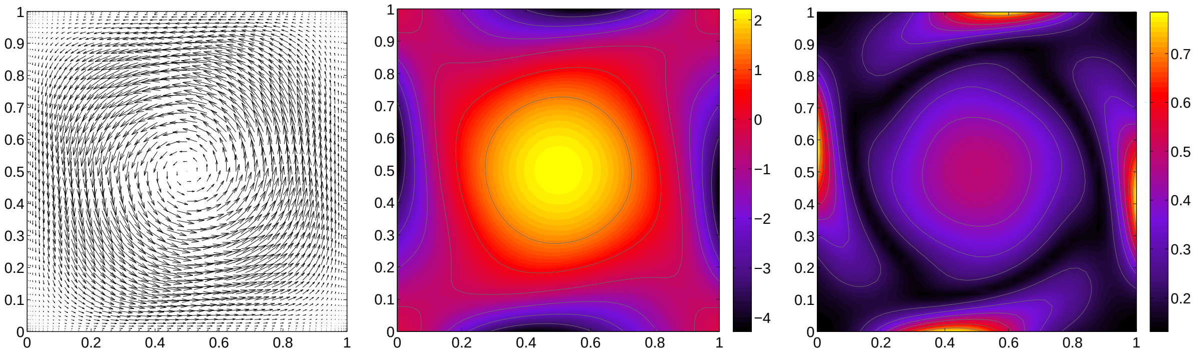
In the second test we consider also the forcing term as uncertain. To this end, we go back to equation (5.21) and take now Φ, the vertical component of the rotational of the force field, as a stationary Gaussian process with unit mean and standard deviation $\sigma_\Phi > 0$, characterized by the two point correlation function

$$C_\Phi(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(\Phi(\mathbf{x}) - \Phi_0)(\Phi(\mathbf{x}') - \Phi_0)\right] = \sigma_\Phi^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{L}\right),$$

where $\Phi_0 = 1$ is the mean of Φ, L its correlation length, and $\|\mathbf{x} - \mathbf{x}'\|$ is the Euclidean norm or $\mathbb{R}^2$. The process admits the Karhunen-Loeve expansion

$$\Phi(\mathbf{x}, \omega) = \Phi_0 + \sum_{i=1}^{\infty} \Phi_i(\mathbf{x}) y_i(\omega),$$

where the y_i are normalized uncorrelated Gaussian variables. Ordering the Karhunen-Loeve modes with decreasing norm $\|\Phi_i\|_{\mathrm{L}^2(D)}$ and truncating the expansion after the N_f-th term results in the



(a) mean velocity field.

(b) mean vorticity field.

(c) standard dev. of the vorticity field.

Figure 5.5: PGD solution of Test 1.

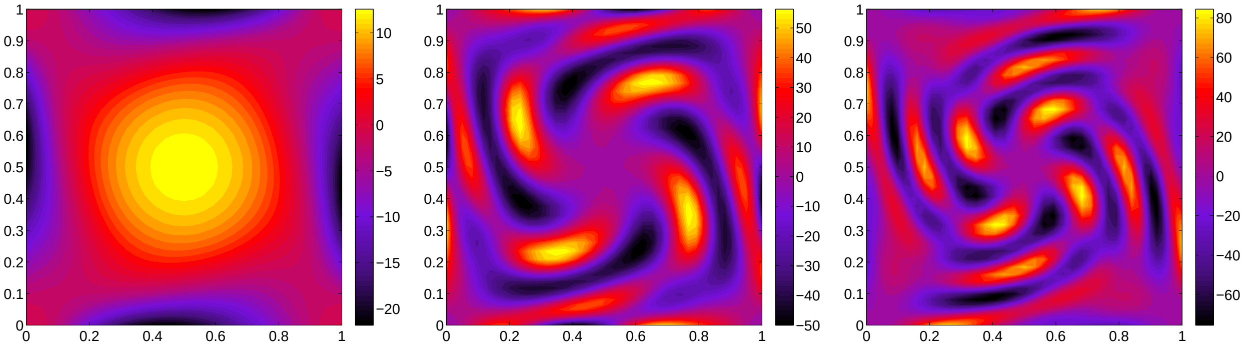


Figure 5.6: Rotational of the KL modes $0, 3, 6$ of the PGD solution of Test 1.

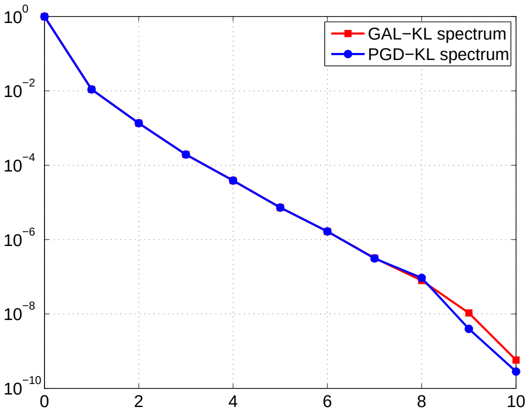


Figure 5.7: comparison of the spectra of the KL decomposition for the Galerkin and PGD solutions of Test 1.

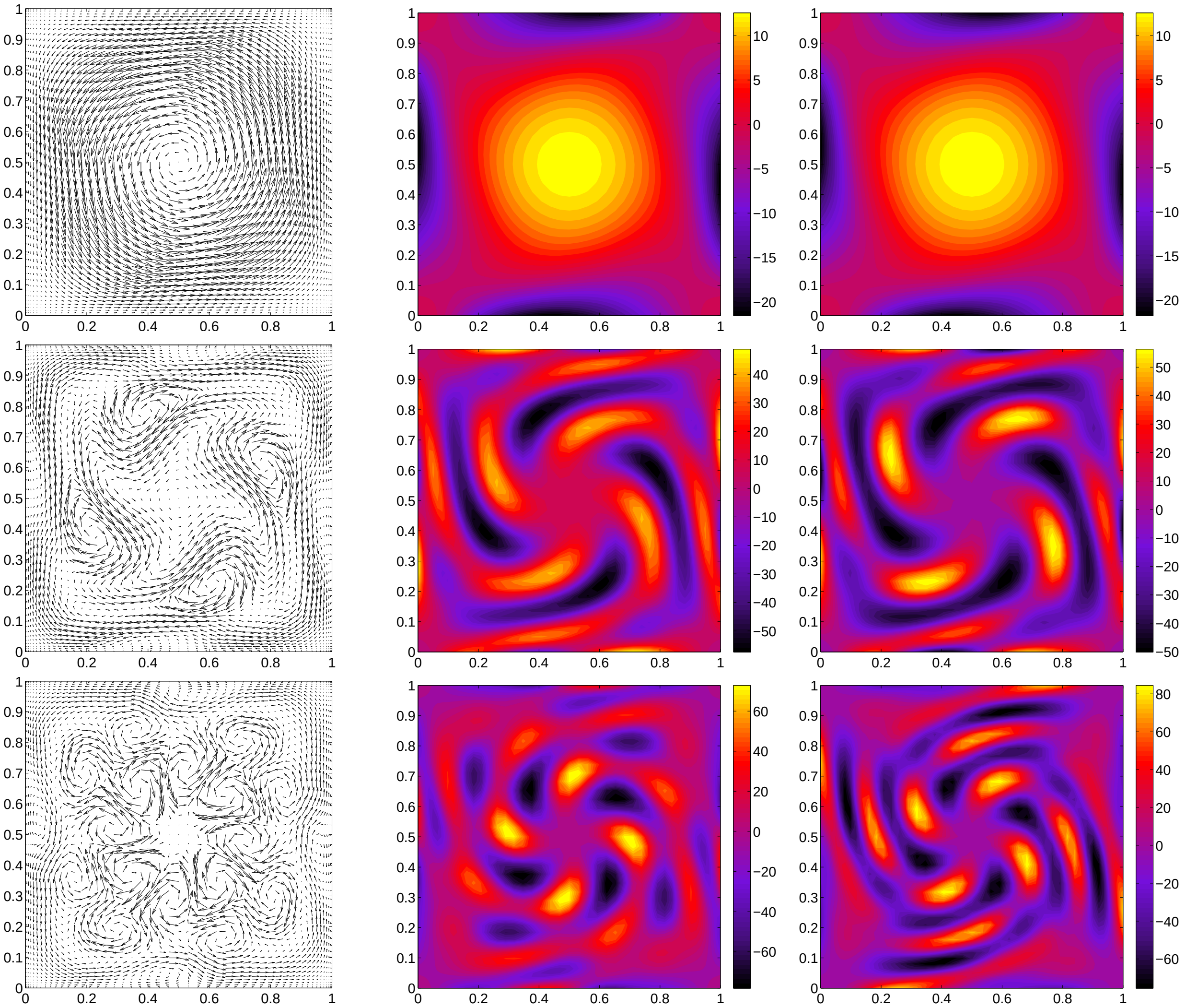
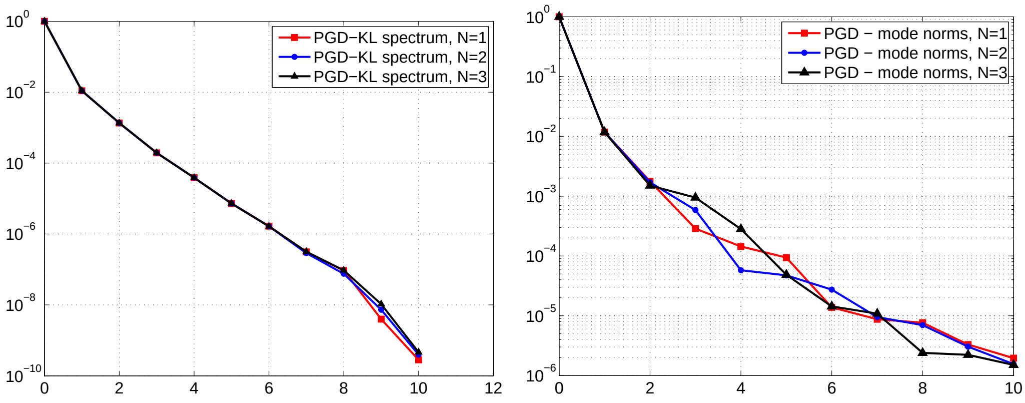


Figure 5.8: PGD modes for $m = 0, 3, 6$ (left), the corresponding rotational (center) and the rotational of the corresponding KL mode of the PGD solution of Test 1 (right).



(a) norms of KL modes.

(b) norms of PGD stochastic modes λ_i

Figure 5.9: Comparison of the PGD solutions of Test 1 with $N_\nu = 1, 2, 3$.

following approximation of the external force field:

$$\mathbf{F}(\mathbf{x},\omega) \approx \mathbf{F}^{N_f}(\mathbf{x},\omega) = \mathbf{F}_0 + \sum_{i=1}^{N_f} y_i(\omega)\,\mathbf{F}_i(\mathbf{x})\,, \quad \mathbf{F}_i(\mathbf{x}) = \boldsymbol{\nabla} \wedge \begin{pmatrix} 0 \\ 0 \\ \mathcal{L}[\Phi_i(\mathbf{x})] \end{pmatrix}. \tag{5.26}$$

We set $L = 1$, $\sigma_\Phi = 0.2$ and $\mathtt{Nu} = 35$. It is well known that as L decreases more and more KL modes are needed to represent accurately the forcing term. However, in this work we are not really concerned about the truncation error that stems from retaining only N_f terms of the expansion, but only to show that the PGD methods can handle such forcing terms in a natural way. We will consider two different choices of N_f, that is $N_f = 3$ or $N_f = 7$.

As for the viscous parameter ν, we consider it again as a lognormal random variable ($N_\nu = 1$), as in equation (5.19). This implies that the solution depends on $N = N_\nu + N_f = 4, 8$ random variables respectively. The discete probability space $\mathbb{P}_M(\Gamma)$ is selected setting $w = 2$, resulting in a span of $M = 15, 45$ multivariate Hermite polynomials; within this setting, we compute the Galerkin solution and the PGD solution with $m = 45$ modes. We will consider three different median values for the viscosity parameter, that is $\overline{\nu}' = 1/10, 1/50, 1/100$.

Figure 5.10 shows mean and standard deviation of the rotational of the PGD solution for the case $N = 8$, $\overline{\nu}' = 1/100$, while Figure 5.11 shows for the same case some of the PGD modes. Finally, Figure 5.12 shows the decay of $\|\mathbf{U}^{(m)} - \mathbf{U}^G\|$ as the number of modes m of the PGD solution increases, for all the cases considered. The PGD approximation is thus seen to converge to the full Galerkin solution: as expected the cases with a smaller N and low viscosity parameter feature a higher convergence rate. in all cases the PGD method gives reasonable approximations of the full Galerkin solution with $m \leq M$: this is more and more evident as the number of random variables increases, see also the results in the next Section where we consider a test case depending on $N = 15$ random variables.

5.5 Residual computation and pressure reconstruction

At this point, it is crucial to devise an error estimator to stop the PGD procedure as soon as the reduced solution is close enough to the exact solution in $\mathbf{H}^1_{0,div}(D) \otimes \mathbb{P}_M(\Gamma)$.

The most natural approach would be a stopping criterion involving the evaluation of the norm of the residual of the Stochastic Navier–Stokes equation (5.16) associated to the m-terms reduced solution $\mathbf{U}^{(m)}$ in the discretized space $\mathbf{H}^1_{0,div}(D) \otimes \mathbb{P}_M(\Gamma)$. The Arnoldi algorithm would then be stopped as soon as such residual becomes lower than a given tolerance in a suitable norm. In practice, computing the residual of the Navier–Stokes equations in their divergence-free formulation (5.16) is not a convenient operation. Therefore, we go back to the weak deterministic Navier–Stokes equations (5.14) and introduce the

Stochastic Velocity-Pressure Navier–Stokes equations.

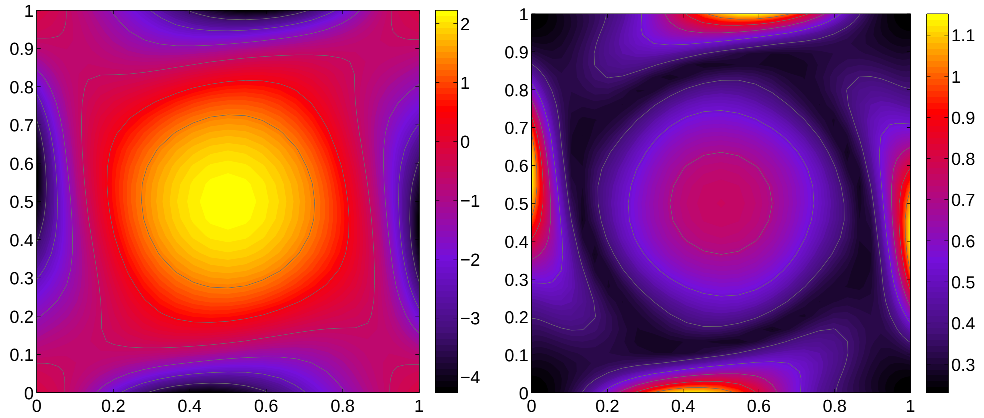
Find $\mathbf{U} \in \mathbf{H}^1_0(D) \otimes \mathbb{P}_M(\Gamma)$, $\mathbf{P} \in \mathrm{L}^2_0(D) \otimes \mathbb{P}_M(\Gamma)$ *such that*

$$\begin{aligned} &C(\mathbf{U},\mathbf{U},\mathbf{V}) + V_\nu(\mathbf{U},\mathbf{V}) + E(\mathbf{P},\mathbf{V}) = B(\mathbf{V}) && \forall \mathbf{V} \in \mathbf{H}^1_0(D) \otimes \mathbb{P}_M(\Gamma), \\ &E(\mathbf{Q},\mathbf{U}) = 0 && \forall \mathbf{Q} \in \mathrm{L}^2_0(D) \otimes \mathbb{P}_M(\Gamma), \end{aligned} \tag{5.27}$$

where $E(\mathbf{Q},\mathbf{V})$ is defined as the expected value of the bilinear form $d(\cdot,\cdot)$ appearing in (5.14),

$$E(\mathbf{Q},\mathbf{V}) = \mathbb{E}\left[d(\mathbf{Q},\mathbf{V})\right].$$

Computing the residual for the velocity-pressure formulation is an affordable task, but at this point the PGD algorithm has not provided us with an approximation of the stochastic pressure yet.



(a) mean vorticity field.

(b) standard deviation of the vorticity field.

Figure 5.10: 45-modes PGD solution of Test 2, $\overline{\nu}' = 1/100$, $N = 8$.

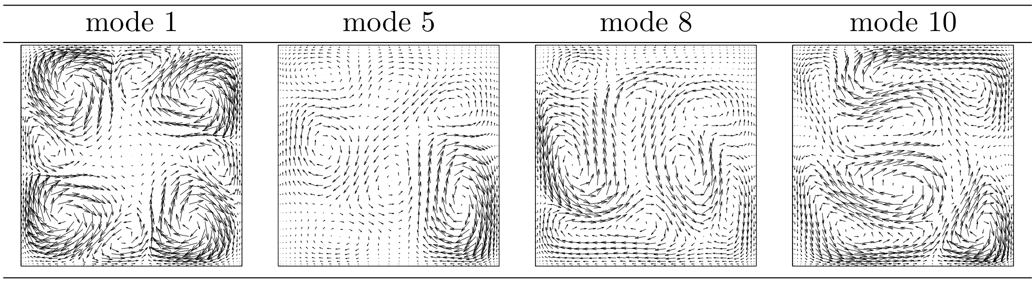
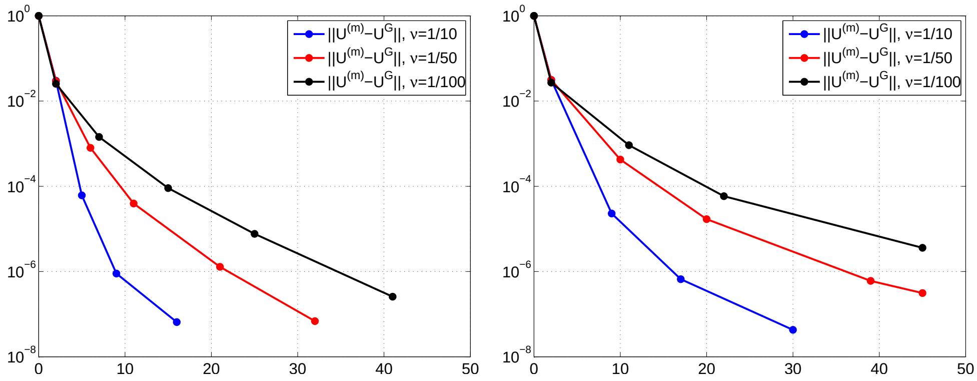


Figure 5.11: PGD modes for the solution of Test 2, $\overline{\nu} = 1/100$, $N = 8$.



(a) $N = 4$.

(b) $N = 8$.

Figure 5.12: convergence of $\|\mathbf{U}^{(m)} - \mathbf{U}^G\|/\|\mathbf{U}^G\|$ with respect to the number of modes m in the PGD solution for $\overline{\nu} = 1/10,\ 1/50,\ 1/100$, Test 2.

Hence, we now introduce a procedure to recover the pressure $\mathbf{P}^{(m)}$ associated to the m-terms PGD solution $\mathbf{U}^{(m)}$.

Computing such approximation will introduce some computational overhead, but one could be interested in an approximation of the pressure anyway. We stress that the notation $\mathbf{P}^{(m)}$ does not refer to an m-terms approximation of $\mathbf{P}$, but to a generic approximation of $\mathbf{P}$ given the m-terms reduced approximation of $\mathbf{U}$.

5.5.a Pressure computation

For easiness of presentation, let us define

$$N(\mathbf{W},\mathbf{V}) = C(\mathbf{W},\mathbf{W},\mathbf{V}) + V_\nu(\mathbf{W},\mathbf{V}) - B(\mathbf{V}), \quad \forall \mathbf{V},\mathbf{W} \in \mathbf{H}_0^1(D)\otimes \mathbb{P}_M(\Gamma), \tag{5.28}$$

and let $\langle \mathbf{V},\mathbf{W}\rangle$ denote the scalar product in $\mathbf{H}_0^1(D)\otimes\mathbb{P}_M(\Gamma)$. Inserting the m-terms PGD velocity $\mathbf{U}^{(m)}$ and the corresponding pressure $\mathbf{P}^{(m)}$ into the Stochastic Velocity-Pressure Navier–Stokes equations (5.27) we have

$$N(\mathbf{U}^{(m)},\mathbf{V}) + E(\mathbf{P}^{(m)},\mathbf{V}) = \left\langle \mathbf{R}^{(m)},\mathbf{V}\right\rangle \qquad \forall \mathbf{V}\in \mathbf{H}_0^1(D)\otimes\mathbb{P}_M(\Gamma), \tag{5.29a}$$

$$E(\mathbf{Q},\mathbf{U}^{(m)}) = 0 \qquad \forall \mathbf{Q}\in \mathrm{L}_0^2(D)\otimes\mathbb{P}_M(\Gamma), \tag{5.29b}$$

where $\mathbf{R}^{(m)}$ denotes the residual of the momentum equation (5.29a), $\mathbf{R}^{(m)}\in\mathbf{H}_0^1(D)\otimes\mathbb{P}_M(\Gamma)$. Note that the continuity equation (5.29b) has no residual; indeed, all the deterministic modes in $\mathbf{U}^{(m)}$ are divergence-free, being solutions of the deterministic problem (5.17). Equation (5.29) states that the residual $\mathbf{R}^{(m)}$ is a function of the pressure $\mathbf{P}^{(m)}$: here we propose to estimate $\mathbf{P}^{(m)}$ as the *minimizer* of $\|\mathbf{R}^{(m)}\|$ in some prescribed norm. To be more computationally oriented, we next derive the problem for $\mathbf{P}^{(m)}$ in the discrete case.

Let us denote with $V_h\subset\mathbf{H}_0^1(D)$ the finite dimensional velocity space, and with $\Pi_h\in\mathrm{L}_0^2(D)$ the finite dimensional pressure space. Upon the introduction of the bases for V_h and Π_h defined in [17] and that will be used in the Section of numerical results, we can identify any element $\mathbf{W}_h\in V_h\otimes\mathbb{P}_M(\Gamma)$ with the coordinates in the respective basis $\widehat{\mathbf{W}}_h(\mathbf{y})\in\mathbb{R}^{\dim(V_h)}\otimes\mathbb{P}_M(\Gamma)$, and similarly any element $\mathbf{Q}_h\in\Pi_h\otimes\mathbb{P}_M(\Gamma)$ with $\widehat{\mathbf{Q}}_h(\mathbf{y})\in\mathbb{R}^{\dim(\Pi_h)}\otimes\mathbb{P}_M(\Gamma)$; in other words, $\widehat{\mathbf{W}}_h(\mathbf{y})$ and $\widehat{\mathbf{Q}}_h(\mathbf{y})$ are vectors whose components are functions of $\mathbf{y}$, belonging to the subspace $\mathbb{P}_M(\Gamma)\subset\mathrm{L}_\rho^2(\Gamma)$. Equation (5.29a) can therefore be recast as a semidiscrete equation in $\mathbb{R}^{\dim(V_h)}\otimes\mathbb{P}_M(\Gamma)$,

$$\widehat{\mathbf{N}}_h^{(m)}(\mathbf{y}) + \mathcal{E}^T\widehat{\mathbf{P}}_h^{(m)}(\mathbf{y}) = \widehat{\mathbf{R}}_h^{(m)}(\mathbf{y}), \tag{5.30}$$

with $\widehat{\mathbf{N}}_h^{(m)}(\mathbf{y}),\widehat{\mathbf{R}}_h(\mathbf{y})\in\mathbb{R}^{\dim(V_h)}\otimes\mathbb{P}_M(\Gamma)$, $\widehat{\mathbf{P}}_h^{(m)}(\mathbf{y})\in\mathbb{R}^{\dim(\Pi_h)}\otimes\mathbb{P}_M(\Gamma)$, and $\mathcal{E}\in\mathbb{R}^{\dim(\Pi_h)\times\dim(V_h)}$ the deterministic discrete divergence operator. Next we define the residual norm as

$$\|\widehat{\mathbf{R}}_h^{(m)}(\mathbf{y})\|^2 = \|\widehat{\mathbf{R}}_h^{(m)}(\mathbf{y})\|^2_{\mathbb{R}^{\dim(V_h)}\otimes\mathbb{P}_M(\Gamma)} = \frac{1}{2}\,\mathbb{E}\left[\|\widehat{\mathbf{R}}_h^{(m)}(\mathbf{y})\|^2_{\mathbb{R}^{\dim(V_h)}}\right],$$

use equation (5.30) and enforce the derivative of $\|\widehat{\mathbf{R}}_h^{(m)}(\mathbf{y})\|$ with respect to $\widehat{\mathbf{P}}_h^{(m)}(\mathbf{y})$ to be zero. Thus we obtain that the pressure minimizing $\|\widehat{\mathbf{R}}_h^{(m)}(\mathbf{y})\|$ is the solution of

$$\mathcal{E}\,\mathcal{E}^T\widehat{\mathbf{P}}_h^{(m)}(\mathbf{y}) = -\mathcal{E}\widehat{\mathbf{N}}_h^{(m)}(\mathbf{y})\,, \tag{5.31}$$

that further needs to be discretized along the stochastic dimension. Note that $\mathcal{E}\,\mathcal{E}^T$ is a deterministic operator, and equation (5.31) is well-posed if V_h and Π_h verify the inf-sup condition. Moreover, computing the gPCE expansion of $\widehat{\mathbf{P}}_h^{(m)}(\mathbf{y})$, that is

$$\widehat{\mathbf{P}}_h^{(m)}(\mathbf{y}) = \sum_{k=1}^{M}\widehat{\mathbf{P}}_{h,k}^{(m)}\mathcal{H}_k(\mathbf{y}),$$

with $\widehat{\mathbf{P}}_{h,k}^{(m)}\in\mathbb{R}^{\dim(\Pi_h)}$ and $\mathcal{H}_k(\mathbf{y})\in\mathbb{P}_M(\Gamma)$ Hermite polynomials, results in a set of M uncoupled problems

$$\mathcal{E}\,\mathcal{E}^T\widehat{\mathbf{P}}_{h,k}^{(m)} = -\mathcal{E}\widehat{\mathbf{N}}_{h,k}^{(m)}\,. \tag{5.32}$$

Note that $\widehat{\mathbf{N}}_{h,k}^{(m)}$ has to be computed, using the projection $\widehat{\mathbf{N}}_{h,k}^{(m)} = \mathbb{E}\,[\widehat{\mathbf{N}}_{h}^{(m)}(\mathbf{y})\mathcal{H}_k(\mathbf{y})]$, since the stochastic vector $\widehat{\mathbf{N}}_{h}^{(m)}(\mathbf{y})$ derives from a non-linear combination of the PGD solution, hence its gPCE expansion is not immediately available.

Even if we can take advantage of the fact that the problems (5.32) are uncoupled by factorizing the operator $\mathcal{E}\mathcal{E}^T$ only once to improve the computational efficiency (e.g. with a LU, ILU or Cholesky factorization), the overall cost may be demanding if the discrete stochastic space $\mathbb{P}_M(\Gamma)$ is large. We have then considered two additional strategies for the computation of the pressure. In the first one, we apply the Arnoldi algorithm illustrated in Section 5.2.e to equation (5.31) to obtain a PDG approximation of the stochastic pressure,

$$\widehat{\mathbf{P}}_{h}^{(m)}(\mathbf{y}) = \sum_{k=1}^{m'} \widehat{\mathbf{P}}_{h,k}^{(m)} \gamma_k(\mathbf{y}), \tag{5.33}$$

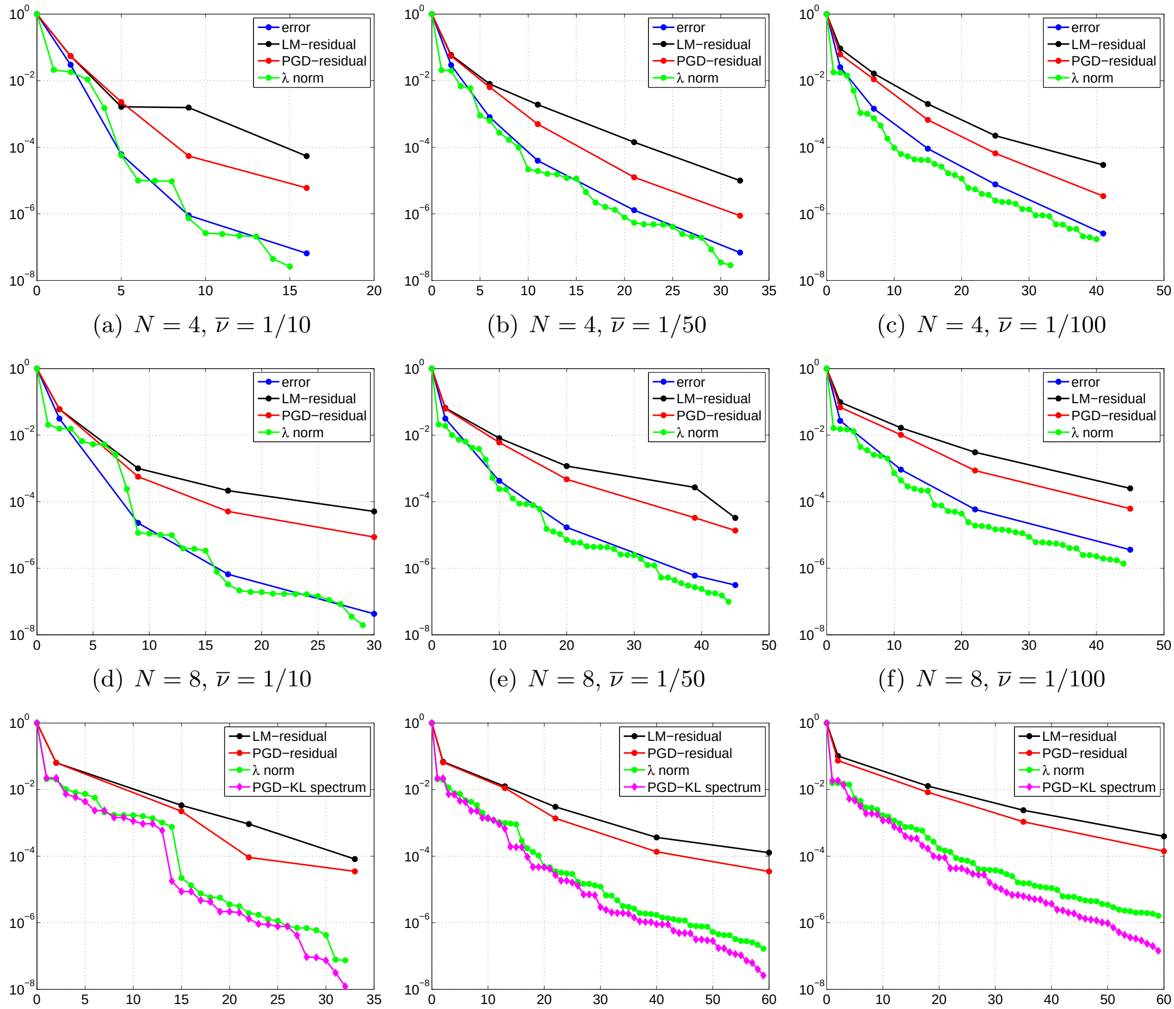
with $\widehat{\mathbf{P}}_{h,k}^{(m)} \in \mathbb{R}^{\dim(\Pi_h)}$ and $\gamma_k(\mathbf{y}) \in \mathbb{P}_M(\Gamma)$ generic functions. Note that the PGD approximation of $\mathbf{P}$ may in general use $m' \neq m$ modes. The second approach we have considered consists in using the Lagrange multipliers obtained from the deterministic problems solved during the computation of the PGD decomposition of $\mathbf{U}^{(m)}$ as deterministic modes for the PGD approximation (5.33) of the pressure. This latter approach allows further savings in terms of computational cost; note that in this case $m = m'$.

5.5.b Numerical results

In the previous section we have proposed two ways of computing an approximation of the pressure field: a "full reduced approach" in which we use a PGD method to compute both the deterministic and the stochastic modes of the decomposition (5.33), and a "partial reduced approach" in which the deterministic modes $\widehat{\mathbf{P}}_{h,k}^{(m)}$ of (5.33) are taken to be the Lagrange multipliers resulting from the solution of the deterministic problem of the Arnoldi iterations. In both cases, the obtained pressure approximation will be then used to compute the residual $\widehat{\mathbf{R}}_{h}^{(m)}(\mathbf{y})$ through equation (5.30), and the norm $\|\widehat{\mathbf{R}}_{h}^{(m)}(\mathbf{y})\|$ will be used as a stopping criterion for the Arnoldi method.

We now aim at assessing the performances of these two stopping criteria, along with a third one, which is the monitoring of $\|\lambda_i\|$, the norm of the stochstic modes of the PGD approximation. This criterion is indeed much simpler, and based on the observation that, since the deterministic modes $\boldsymbol{u}_i$ of the Arnoldi method are normalized, whenever λ_i is small, we are "adding nothing" to the reduced approximation. Such criterion may be reasonable whenever one is not at all interested in pressure recostruction.

We consider again the setting proposed for the second test presented in the previous section, with both viscosity and forcing terms modeled as random quantitites (see section 5.4.b). We consider the cases $N = 4, 8$ and we also add a new case, where we set $N = 15$ and consider a truncation with $w = 3$, resulting in a span of $M = 816$ stochastic polynomials. The convergence of the proposed quantities for Test 2 is shown in Figures 5.13, 5.14. 5.15. The residual computed by recycling the Lagrange Multipliers is slightly worse than the one computed after having reconstructed the pressure with a PGD approach, and they both overestimate the error by 1-2 orders of magnitude, hence representing a quite restrictive criterion for the convergence ot the method. On the other hand, the norms of the λ_i appear closer to the true error, but slightly underestimating it, hence representing an "optimistic" criterion.



(a) $N = 4$, $\overline{\nu} = 1/10$ (b) $N = 4$, $\overline{\nu} = 1/50$ (c) $N = 4$, $\overline{\nu} = 1/100$

(d) $N = 8$, $\overline{\nu} = 1/10$ (e) $N = 8$, $\overline{\nu} = 1/50$ (f) $N = 8$, $\overline{\nu} = 1/100$

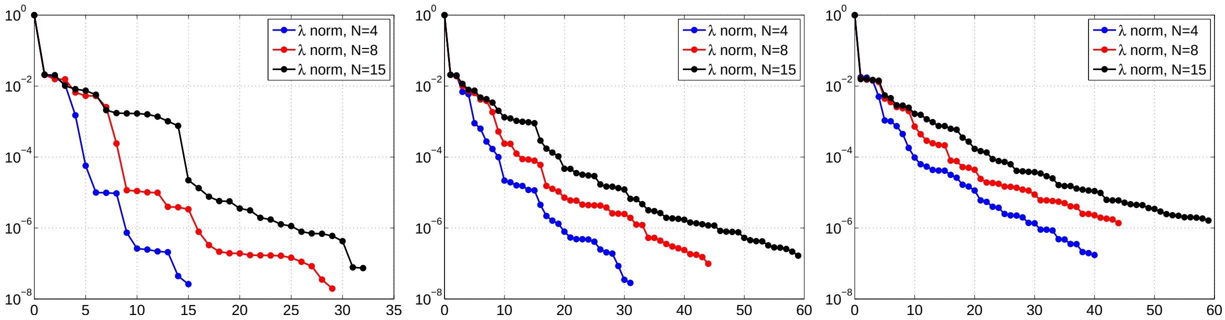
(g) $N = 15$, $\overline{\nu} = 1/10$ (h) $N = 15$, $\overline{\nu} = 1/50$ (i) $N = 15$, $\overline{\nu} = 1/100$

Figure 5.13: convergence of the quantities proposed as stopping criterion for the PGD method with respect to the number of modes m. i) “error” denotes the normalized PGD-Galerkin error $\|\mathbf{U}^{(m)} - \mathbf{U}^G\|/\|\mathbf{U}^G\|$; ii) “LM- residual” denotes the normalized norm of residual $\|\widehat{\mathbf{R}}_h^{(m)}(\mathbf{y})\|/\|\widehat{\mathbf{R}}_h^{(0)}(\mathbf{y})\|$, the residual being computed using the Lagrange Multipliers as deterministic modes for the pressure; iii) “PGD- residual” denotes the normalized norm of residual $\|\widehat{\mathbf{R}}_h^{(m)}(\mathbf{y})\|/\|\widehat{\mathbf{R}}_h^{(0)}(\mathbf{y})\|$, the residual being computed using the pressure reconstructed with a PGD approach; iv) “λ norm” denotes the normalized norm of λ_i, that is $\|\lambda_i\|/\sqrt{\sum_i \|\lambda_i\|}$.

5.6 Conclusions

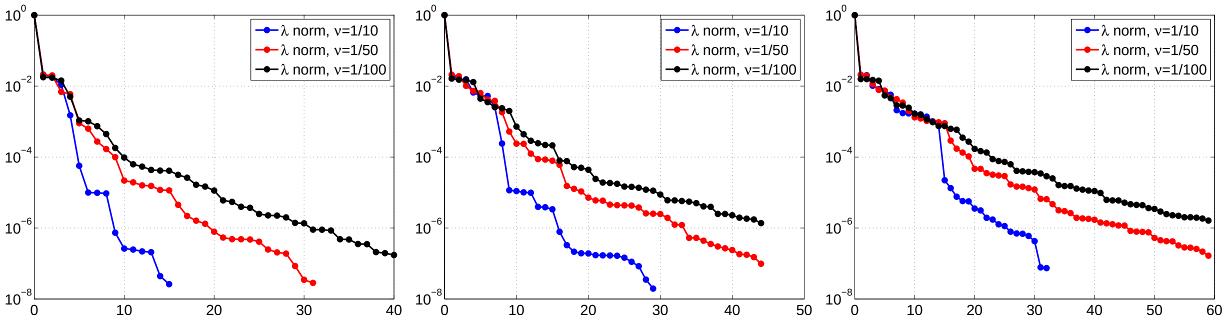
In this Chapter we have investigated the resolution of the steady-state Navier–Stokes equations with uncertain parameters through a *PGD* procedure; in particular, we have employed the Arnoldi method in our numerical simulations, but also the Power method (possibly with update) could be considered as well.

We recall that the cost of the *PGD* -Arnoldi method consists of one deterministic solver call per mode retained in the *PGD* expansion, plus the same number of the stochastic problem solver calls and a few calls to the update problem solver (usually $5-6$ calls in the test we have performed), which



(a) $\nu = 1/10$ (b) $\nu = 1/50$ (c) $\nu = 1/100$

Figure 5.14: Comparing the convergence of $\|\lambda\|$ for different N fixed ν.



(a) $N = 4$ (b) $N = 8$ (c) $N = 15$

Figure 5.15: Comparing the convergence of $\|\lambda\|$ for different N fixed ν.

however is not a negligible cost, since the update problem is a system of quadratic equations, whose dimension is $m \times M$. Additional costs may be introduced by a pressure reconstruction procedure.

This cost should be compared to the cost of the "Optimal Sets" approach for the Galerkin method detailed in Chapter 3, since both methods are seen as improvements with respect to the standard Stochastic Galerkin method. A comparison is however not immediate, since the cost of the "Optimal Sets" approach can be estimated in terms of calls to a deterministic solver solely, and will depend both on the sharpness of the bound used to compute the optimal set and on the efficiency of the solving technique (e.g. Preconditioned Conjugated Gradient).

A remarkable advantage of the *PGD* procedure with respect to the standard Galerkin techniques relies in the coding effort required: it is indeed possible in this framework to reuse any existing Navier–Stokes solver to solve the deterministic steps of the Arnoldi method, taking care of the modifications of the convective velocity, the viscosity parameter and the forcing term. The stochastic and update problems can also be solved with available software, since they amount to systems of quadratic equations. However, care has to be taken in the recostruction of a reduced pressure: the mathematical formulation of this problem is non-trivial, and we have addressed this topic only for the discrete problem, proposing different approaches.

The convergence of the *PGD* velocity to the full Galerkin solution has been investigated with three different numerical settings, resulting in parametrizations with $N = 4, 8, 15$ random variables. In all the considered cases, the *PGD* is able to provide reasonable approximations of the full Galerkin solution (10^{-4}) with a limited number of modes, thus with a (much) smaller compuational cost compared to the solution of the full Galerkin problem.

Chapter 6

Global sensitivity analysis using sparse grids: an application to a geochemical compaction model

This Chapter is derived from the paper by L. Formaggia, A. Guadagnini, I. Imperiali, G. Porta, M. Riva, A. Scotti, L. Tamellini, *A numerical model for the geological compaction of sedimentary basins with sensitivity analysis*, currently in preparation. Our contribution has been to take care of the uncertainty quantification mathematical framework and to provide the code for the uncertainty quantification analysis. In particular, we have developed an algorithm to convert a sparse grid approximation to a gPCE expansion, see Section 6.2 for details.

Therefore, compared to the original paper, in this Chapter we have modified the order of the exposition, and rewritten a significative part of the text, to put in evidence the part of interest for this thesis, that is the uncertainty quantification analysis rather than the deterministic problem itself. The mathematical model and its discretization are thus here only briefly summarized, and the literature survey on the geological compaction has been reduced.

6.1 Introduction

In the previous Chapters we have mostly focused on the theoretical aspects related to the construction of a surrogate polynomial model for the approximation of the solution of PDEs with stochastic coefficients.

In this Chapter we focus instead on the practical application of such surrogate model for uncertainty quantification analysis. In particular, we will concentrate on a global sensitivity analysis, that is the evaluation of the influence of each random parameter and of each mixed effect (resulting from the combination of random parameters) to the uncertainty of the outcomes of the system. A review of sensitivity analysis practices can be found in [92], where it is proved that a global sensitivity analysis is required unless the underlying model is linear.

Among the possible probabilistic global sensitivity analysis techniques, e.g. ANOVA techniques [53] and design of experiments [68], we consider in this Chapter the Sobol' sensitivity indices ([96, 1, 92, 99]). These can be indeed easily computed by exploiting the orthogonality properties of the generalized polynomial chaos expansion (gPCE) of the solution, see e.g. [60, 99]. A comparison between global sensitivity analysis performed with Sobol'indices and ANOVA techniques can be found e.g. in [1].

In this Chapter we will perform a global sensitivity analysis on a model for the geochemical compaction of sediments, that is the process that transforms sea sediments into rocks. The process is driven by both mechanical stresses, like the weight of the upper layers pressing the lower layers, and by chemical reactions, like quartz precipitation. The set of equations for the model thus includes force and energy balance, the Darcy equation for the expulsion of water from sediments, a number of constitutive laws that model physical quantities like porosity, permeability, thermal conductivity etc., and the chemical reactions occurring in the sedimentary basin. Most of the scalar parameters of these equations are unknown and modeled as uniform random variables (in our tests we have considered $N = 7$ random parameters). The goal of the global sensitivity analysis is to determine what phenomena (mechanical and/or chemical) affects the most the transformation process. The main quantity of interest will be therefore the porosity of the strata of the basin, that is the free space among rock grains.

Given the complexity of the model for the deterministic problem at hand, a Galerkin approach for the computation of the gPCE expansion would result in a system of equation whose matrix is extremely difficult to assemble. We prefer therefore a collocation approach on a sparse grid. The resulting polynomial approximation will be then converted onto the Legendre orthonormal basis (gPCE expansion), exploiting the equivalence Theorem stated in Chapter 2. This approach to compute a gPCE expansion is quite new and only few similar works are available in literature; we mention e.g. [22]. It is more robust than simply using a sparse grid quadrature to compute the gPCE coefficients. Indeed it can be easily shown that, with the proposed approach, the error on the approximation of all the gPCE coefficients is controlled by the sparse grid approximation error, whereas the direct computation of the gPCE coefficients with the same sparse grid quadrature may lead to inaccurate estimates of the high-order terms.

The rest of this Chapter is organized as follows. In Section 6.2 we detail the Sobol'indices computation through the conversion of a sparse grid into a gPCE expansion. In Section 6.3 we briefly summarize the deterministic problem of interest. Results of the sensitivity analysis performed are detailed in Section 6.4.

6.2 Sparse grids-driven uncertainty quantification analysis

Consider a generic deterministic problem depending on a set of N independent random parameters, $\mathbf{y} = (y_1, y_2, \dots, y_N)^T$. Let each parameter y_i take value in Γ_i, and denote with ρ_i its probability density function. Thus $\mathbf{y}$ takes values in $\Gamma = \Gamma_1 \times \Gamma_2 \ldots \times \Gamma_N$, and the joint probability function is simply the product of each marginal probability density function, $\rho_\Gamma(\mathbf{y}) = \prod_{i=1}^N \rho_{\Gamma_i}(y_i)$.

Any scalar quantity of interest related to the deterministic problem can then be seen as a random function, $\mathcal{Q} = \mathcal{Q}(\mathbf{y})$. As mentioned in the Introduction, we are intersted in performing an uncertainty quantification analysis for $\mathcal{Q}$, and in particular we aim at obtaining a complete description of how the total variance of the quantity of interest can be attributed to each random variable y_i and to each mixed effect $y_{i_1} y_{i_2} \ldots y_{i_s}$ (global sensitivity analyisis).

6.2.a Sobol' indices

We now introduce the Sobol' indices to perform a global sensitivity analysis. Such indices are interesting since they can be computed from the gPCE expansion of $\mathcal{Q}(\mathbf{y})$ and do not assume any linearity in the model considered (see e.g. [92]).

Following [60, 96, 99], we first introduce the Hoeffding/Sobol' decomposition of a generic function f depending on N independent random variables. Let $\{i_1, i_2, \dots, i_s\} \subseteq \{1, \dots, N\}$ be a set of indices; we denote with $\mathbf{y}_{\sim\{i_1,i_2,\dots,i_s\}}$ the set of all random variables but $y_{i_1}, y_{i_2}, \ldots y_{i_s}$, and similarly we let $\Gamma_{\sim\{i_1,i_2,\dots,i_s\}}$ be the cartesian product of all the domains but $\Gamma_{i_1}, \Gamma_{i_2}, \ldots \Gamma_{i_s}$ and $\rho_{\Gamma_{\sim\{i_1,i_2,\dots,i_s\}}}$

the product of all the probability density functions but $\rho_{i_1}, \rho_{i_2}, \dots \rho_{i_s}$. The Hoeffding/Sobol' decomposition is then defined as

$$
\begin{aligned}
f(\mathbf{y}) &= f_0 + \sum_{i=1}^{N} f_i(y_i) + \sum_{i=1}^{N} \sum_{j>i}^{N} f_{ij}(y_i, y_j) \ + \ldots + f_{1,2,\ldots,N}(y_1, y_2, \ldots, y_N), \qquad (6.1)\\
f_0 &= \int_\Gamma f(\mathbf{y}) \rho_\Gamma(\mathbf{y}) d\mathbf{y},\\
f_{\{i_1,i_2,\ldots,i_s\}} &= \int_{\Gamma_{\sim\{i_1,i_2,\ldots,i_s\}}} f(\mathbf{y}) \rho_{\Gamma_{\sim\{i_1,i_2,\ldots,i_s\}}}(\mathbf{y}) d\mathbf{y}_{\sim\{i_1,i_2,\ldots,i_s\}} \ - \sum_{\mathcal{S} \subset \{\{i_1,i_2,\ldots,i_s\}\}} f_{\mathcal{S}}.
\end{aligned}
$$

For example

$$
\begin{aligned}
f_i(y_i) &= \int_{\Gamma_{\sim i}} f(\mathbf{y}) \rho_{\Gamma_{\sim i}}(\mathbf{y}) d\mathbf{y}_{\sim i} - f_0,\\
f_{i,j}(y_i, y_j) &= \int_{\Gamma_{\sim i,j}} f(\mathbf{y}) \rho_{\Gamma_{\sim i,j}}(\mathbf{y}) d\mathbf{y}_{\sim i,j} - f_i(y_i) - f_j(y_j) - f_0.
\end{aligned}
$$

Note that f_0 is the mean of f, and we have the following lemma, whose proof is immediate.

Lemma 6.1. *Given an index $i^* \in \{i_1, i_2, \ldots, i_s\}$, it holds*

$$
\int_{\Gamma_i^*} f_{\{i_1,i_2,\ldots,i_s\}} dy_{i^*} = 0. \qquad (6.2)
$$

Thanks to Lemma 6.1, it is then possible to prove the following Lemma, that states the orthogonality of all terms in (6.1), see [96].

Lemma 6.2. *The terms in* (6.1) *are mutually orthogonal, i.e.*

$$
\{i_1, i_2, \ldots, i_s\} \neq \{j_1, j_2, \ldots, j_r\} \ \Rightarrow \ \int_\Gamma f_{\{i_1,i_2,\ldots,i_s\}} f_{\{j_1,j_2,\ldots,j_r\}} \rho_\Gamma(\mathbf{y}) d\mathbf{y} = 0\,. \qquad (6.3)
$$

Next, we define the Sobol' index $S_{\{i_1,i_2,\ldots,i_s\}}$ corresponding to the mixed effect $y_{i_1} y_{i_2} \ldots y_{i_s}$ as

$$
S_{\{i_1,\ldots,i_s\}} = \frac{1}{\mathbb{V}\mathrm{ar}\,[f]} \int_{\Gamma_{\{i_1,\ldots,i_s\}}} f^2_{\{i_1,\ldots,i_s\}}(y_{i_1} \ldots y_{i_s}) \rho_{\Gamma_{\{i_1,\ldots,i_s\}}}(\mathbf{y}) dy_{i_1} \ldots y_{i_s} \qquad (6.4)
$$

with $\Gamma_{\{i_1,\ldots,i_s\}} = \Gamma_{i_1} \times \ldots \times \Gamma_{i_s}$ and $\rho_{\Gamma_{\{i_1,\ldots,i_s\}}}(\mathbf{y})$ the corresponding probability measure. Using the Sobol indices we can then compute a variance decomposition equivalent to the classical ANOVA one, as stated in the next Lemma.

Lemma 6.3. *It holds*

$$
1 = \sum_{i=1}^{N} S_i + \sum_{i=1}^{N} \sum_{j>i}^{N} S_{ij} \ + \ldots + S_{1,2,\ldots,N}\,. \qquad (6.5)
$$

<u>Proof.</u> Integrate the square of (6.1) and exploit the orthogonality property (6.3). □

From this lemma we see that the term $S_{\{i_1,i_2,\ldots,i_s\}}$ represents the percentual contribution of the mixed effect $y_{i_1} y_{i_2} \ldots y_{i_s}$ to the total variance of f. One can also introduce the total index S_i^T describing the total variability due to the i-th random parameter, as the sum over all mixed effects including y_i,

$$
S_i^T = \sum_{\mathcal{S}_i} S_{\{i_1,i_2,\ldots,i_s\}}, \qquad (6.6)
$$

where the summation is taken over the set $\mathcal{S}_i$ of all multi-indices $\{i_1, i_2, \dots, i_s\}$ of any length such that at least one component is i.

The Sobol' indices can be easily computed, without performing any numerical quadrature to approximate the coefficients $f_{\{i_1,i_2,\dots,i_s\}}$ in (6.1), starting from the generalized Polinomial Chaos expansion of f. To this end, consider the family of $\rho_\Gamma(\mathbf{y})d\mathbf{y}$-orthogonal polynomials $\mathcal{L}_\mathbf{p}(\mathbf{y})$, where $\mathbf{p}$ is a multiindex in $\mathbb{N}^N$ and as usual $\mathcal{L}_\mathbf{p}(\mathbf{y}) = \prod_{n=1}^N \mathcal{L}_{n,p_n}(y_n)$, $\mathcal{L}_{n,p_n}(y_n)$ being the family of $\rho_{\Gamma_n}(y_n)dy_n$-orthogonal monodimensional polynomials (see Chapters 2, 3). Note that whenever $\mathbf{p}$ is such that $p_n = 0$, $\mathcal{L}_\mathbf{p}(\mathbf{y})$ is actually independent of y_n, since it holds $\mathcal{L}_0(y_n) = 1$. As a consequence, we can reorder the classical gPCE expansion

$$f(\mathbf{y}) = \sum_{\mathbf{p}\in\mathbb{N}^N} \alpha_\mathbf{p}\mathcal{L}_\mathbf{p}(\mathbf{y}), \quad \alpha_\mathbf{p} = \int_\Gamma f(\mathbf{y})\mathcal{L}_\mathbf{p}(\mathbf{y})\rho_\Gamma(\mathbf{y})d\mathbf{y}, \tag{6.7}$$

to make it equivalent to (6.1),

$$f(\mathbf{y}) = \alpha_\mathbf{0} + \sum_{i=1}^N \sum_{\mathbf{p}\in\mathcal{P}_i} \alpha_\mathbf{p}\mathcal{L}_\mathbf{p}(\mathbf{y}) + \sum_{i=1}^N \sum_{j>i}^N \sum_{\mathbf{p}\in\mathcal{P}_{i,j}} \alpha_\mathbf{p}\mathcal{L}_\mathbf{p}(\mathbf{y}) + \dots, \tag{6.8}$$

where $\mathcal{P}_i$ contains all the multiindices such that only the i-th component is different from 0, $\mathcal{P}_i = \{\mathbf{p} \in \mathbb{N}^N : p_i \neq 0, p_k = 0 \text{ for } k \neq i\}$. Similarly $\mathcal{P}_{\{i_1,i_2,\dots,i_s\}} = \{\mathbf{p} \in \mathbb{N}^N : p_{\{i_1,i_2,\dots,i_s\}} \neq 0,\, p_{\sim\{i_1,i_2,\dots,i_s\}} = 0\}$.

Exploiting (6.8) and the orthonormality of $\mathcal{L}_\mathbf{p}(\mathbf{y})$, we have the following equivalence between gPCE coefficients and Sobol' indices:

$$S_{\{i_1,i_2,\dots,i_s\}} = \sum_{\mathbf{p}\in\mathcal{P}_{\{i_1,i_2,\dots,i_s\}}} \frac{\alpha_\mathbf{p}^2}{\mathbb{V}\mathrm{ar}\,[f]}, \quad \mathbb{V}\mathrm{ar}\,[f] = \sum_{\mathbf{p}\in\mathbb{N}^N} \alpha_\mathbf{p}^2\,. \tag{6.9}$$

6.2.b Sparse grids computation of gPCE

Equation (6.9) provides a fast way to compute Sobol' indices once the gPCE expansion (6.8) for $f(\mathbf{y})$ has been determined. However, computing the coefficients $\alpha_\mathbf{p}$ by Galerkin projection as done in Chapters 2 and 3 may not be feasible if the deterministic problem has a very complex structure and is non-linear (as in the case considered in this Chapter, see next Sections).

To circumvent this problem, a sparse grid approach may be used. A first solution simply entails the computation of approximated gPCE coefficients (6.7) using a sparse grid. A second option, which we follow here, is to compute a sparse grid approximation of the state variables in a given polynomial space $\mathbb{P}$ (see Section 2.3.b), and then to re-express such polynomial approximation in terms of orthogonal polynomials, thus obtaining an approximation of the truncation in $\mathbb{P}$ of the gPCE expansion. Once obtained such gPCE, the Sobol' indices are computed according to equations (6.9). With this latter procedure, the error on every gPCE coefficient in $\mathbb{P}$ is controlled by the sparse grid approximation error, as shown in the following lemma, while the former technique yields approximation of the gPCE coefficients which get worse as $|\mathbf{p}|$ increases (see e.g. the numerical results shown at the end of this section). A similar approach can be found in [22] where however the analysis is confined to standard Smolyak sparse grids and there is no explicit reference to the properties of the underlying polynomial space.

Lemma 6.4. *Let $\alpha_\mathbf{p}$ be a coefficient of the gPCE expansion (6.7) of f truncated to $\mathbb{P}$, and let $\alpha_\mathbf{p}^*$ its approximation obtained by converting the sparse grid approximation of f onto the orthonormal basis. Then*

$$|\alpha_\mathbf{p} - \alpha_\mathbf{p}^*| \leq \|f - f_{SG}\|_{L^2_\rho(\Gamma)}\,.$$

Proof.

$$|\alpha_{\mathbf{p}} - \alpha^*_{\mathbf{p}}| = \left| \int_\Gamma f(\mathbf{y}) \mathcal{L}_{\mathbf{p}} \rho_\Gamma(\mathbf{y}) d\mathbf{y} - \int_\Gamma f_{SG}(\mathbf{y}) \mathcal{L}_{\mathbf{p}} \rho_\Gamma(\mathbf{y}) d\mathbf{y} \right| = \left| \int_\Gamma \left(f(\mathbf{y}) - f_{SG}(\mathbf{y}) \right) \mathcal{L}_{\mathbf{p}} \rho_\Gamma(\mathbf{y}) d\mathbf{y} \right|$$
$$\leq \|f - f_{SG}\|_{L^2_\rho(\Gamma)} \|\mathcal{L}_{\mathbf{p}}\|_{L^2_\rho(\Gamma)} = \|f - f_{SG}\|_{L^2_\rho(\Gamma)}$$

where we have exploited the properties of the scalar product over $L^2_\rho(\Gamma)$ and the fact that $\mathcal{L}_{\mathbf{p}}$ are ρ_Γ-orthonormal. □

We refer to Chapters 2, 3 for the sparse grid construction. Yet, to perform the conversion from sparse grid approximation to gPCE , we need to introduce some additional notation. For a given interpolation level i_n in the n-th direction, let $\mathcal{H}_n^{m(i_n)} = \{y^{i_n}_{n,1}, y^{i_n}_{n,2}, \ldots, y^{i_n}_{n,m(i_n)}\} \subset \Gamma_n$ be a set of $m(i_n)$ interpolation points, and let $\mathscr{L}^{i_n}_{n,k}(y_n)$ be the set of Lagrangian polynomials over such set of points, $k = 1, \ldots, m(i_n)$, so that the corresponding monodimensional interpolant operator is

$$\mathcal{U}_n^{m(i_n)}[g](y_n) = \sum_{k=1}^{m(i_n)} g(y^{i_n}_{n,k}) \mathscr{L}^{i_n}_{n,k}(y_n), \qquad g \in \mathcal{C}^0(\Gamma_n).$$

Taking Cartesian products of the sets $\mathcal{H}_n^{m(i_n)}$, $i = 1, \ldots, N$, we can build a tensor grid, $\mathcal{H}_{\mathbf{i}} = \bigotimes_{n=1}^N \mathcal{H}_n^{m(i_n)}$, that has $M_{\mathbf{i}} = \prod_{n=1}^N m(i_n)$ points. Each point of the tensor grid can be addressed by a multi-index $\mathbf{k} \in \mathbb{N}^N$, $\mathbf{y}^{\mathbf{i}}_{\mathbf{k}} = (y^{i_1}_{1,k_1}, y^{i_2}_{2,k_2}, \ldots, y^{i_N}_{N,k_N})$, $1 \leq k_n \leq m(i_n)$, and the corresponding tensor Lagrange polynomial can be computed as $\mathscr{L}^{\mathbf{i}}_{\mathbf{k}}(\mathbf{y}) = \prod_{n=1}^N \mathscr{L}^{i_n}_{n,k_n}(y_n)$, where the superscript $\mathbf{i}$ refers to the tensor grid the Lagrange polynomial is built on. The tensor grid interpolant for f is then defined as

$$f_{TG,\mathbf{i}}(\mathbf{y}) = \bigotimes_{n=1}^N \mathcal{U}_n^{m(i_n)}[f](\mathbf{y}) = \sum_{\mathbf{y}^{\mathbf{i}}_{\mathbf{k}} \in \mathcal{H}_{\mathbf{i}}} f(\mathbf{y}^{\mathbf{i}}_{\mathbf{k}}) \mathscr{L}^{\mathbf{i}}_{\mathbf{k}}(\mathbf{y}) \tag{6.10}$$

Recalling the definitions of detail operator and hierarchical surplus operator (see Chapters 2,3),

$$\Delta_n^{i_n}[f] = \mathcal{U}_n^{m(i_n)}[f] - \mathcal{U}_n^{m(i_n - 1)}[f], \qquad \mathbf{\Delta}^{\mathbf{i}}[f] = \bigotimes_{n=1}^N \Delta_n^{i_n}[f], \tag{6.11}$$

we can put in evidence the sparse grid approximation on a given set $\mathcal{I}$ as a linear combination of tensor Lagrange polynomials

$$f_{SG} = \sum_{\mathbf{i} \in \mathcal{I}} \mathbf{\Delta}^{\mathbf{i}}[f] = \sum_{\mathbf{i} \in \mathcal{I}} c_{\mathbf{i}} \bigotimes_{n=1}^N \mathcal{U}_n^{m(i_n)}[f](\mathbf{y}) = \sum_{\mathbf{i} \in \mathcal{I}} c_{\mathbf{i}} \sum_{\mathbf{y}^{\mathbf{i}}_{\mathbf{k}} \in \mathcal{H}_{\mathbf{i}}} f(\mathbf{y}^{\mathbf{i}}_{\mathbf{k}}) \mathscr{L}^{\mathbf{i}}_{\mathbf{k}}(\mathbf{y}), \tag{6.12}$$

see Chapters 2, 3 for details on the possible choices of the set of hierarchical surpluses $\mathcal{I}$, and equation (3.27) for the value of $c_{\mathbf{i}}$.

In particular, in Chapter 2 we have shown how to choose the set of indices $\mathcal{I}$ so that the resulting sparse grid approximation belongs to a given polynomial space $\mathbb{P}$ (see Section 2.3.b for a precise statement of this fact and a table of equivalences). Once a suitable space $\mathbb{P}$ has been chosen, it will be enough to re-express the sum of lagrangian polynomials in the sparse grid representation (6.12) as a sum of the $\rho(\mathbf{y})d\mathbf{y}$-orthogonal polynomials spanning $\mathbb{P}$ to obtain a gPCE representation (6.8) of the solution, and hence to compute the Sobol' indices, avoiding any explicit numerical quadrature.

Consider now the $\mathbf{i}$-th tensor grid, $\mathcal{H}_{\mathbf{i}}$. The corresponding Lagrangian interpolant $f_{TG,\mathbf{i}}(\mathbf{y})$ (6.10) is a sum of multidimensional Lagrange polynomials over $m(i_1) \times m(i_2) \times \ldots \times m(i_N)$ points, and therefore their polynomial degree is $(m(i_1)-1)\times(m(i_2)-1)\times\ldots\times(m(i_N)-1)$. Thus, such sum can be recast as a linear combination of all the Legendre polynomials whose maximum degree along direction n does not exceed $m(i_n)-1$, that is

$$f_{TG,\mathbf{i}}(\mathbf{y})=\sum_{\mathbf{p}\in\mathcal{C}_\mathbf{i}}\beta_\mathbf{p}\mathcal{L}_\mathbf{p}(\mathbf{y}),\qquad \mathcal{C}_\mathbf{i}=\{\mathbf{p}\in\mathbb{N}^N: p_n\le m(i_n)-1,\, n=1,\ldots,N\}.$$

The coefficients $\beta_\mathbf{p}$ of the linear expansion can be easily computed enforcing $\beta_\mathbf{p}$ to satisfy the following set of equations

$$\sum_{\mathbf{p}\in\mathcal{C}_\mathbf{i}}\beta_\mathbf{p}\mathcal{L}_\mathbf{p}(\mathbf{y_j})=f_{TG,\mathbf{i}}(\mathbf{y_j}),\quad \forall\mathbf{y_j}\in\mathcal{H}_\mathbf{i}. \tag{6.13}$$

which amounts at solving the linear system $Q\boldsymbol{\beta}=\boldsymbol{f}$, defined (with an abuse of notation) as $Q_{\mathbf{j},\mathbf{p}}=\mathcal{L}_\mathbf{p}(\mathbf{y_j})$, $\boldsymbol{\beta}=[\beta_\mathbf{1},\beta_\mathbf{2},\ldots]$, $\boldsymbol{f}=[f(\mathbf{y_1}),f(\mathbf{y_2}),\ldots]$. Converting the sparse grid into a gPCE expansion therefore entails solving as many linear systems $Q\boldsymbol{\beta}=\boldsymbol{f}$ as the number of tensor grids with non-zero coefficient $c_\mathbf{i}$ in the sparse grid, and then collecting coefficients $\beta_\mathbf{p}$ for the same $\mathbf{p}$ coming from different tensor grids. Note that the properties of Q depend on the choice of the interpolation points $\mathcal{H}_n^{m(i_n)}$ used in each direction. In particular, if the interpolation points are Gaussian, it holds that $Q^TWQ=I$, where W denotes the diagonal matrix containing the quadrature weights, and I is the identity matrix. Hence, $\widetilde{Q}=W^{1/2}Q$ is orthogonal and therefore $\boldsymbol{\beta}=Q^TW\boldsymbol{f}$. Indeed, it holds

$$\boldsymbol{\beta}=Q^{-1}\boldsymbol{f}=Q^{-1}W^{-1/2}W^{1/2}\boldsymbol{f}=(W^{1/2}Q)^TW^{1/2}\boldsymbol{f}=Q^TW\boldsymbol{f}.$$

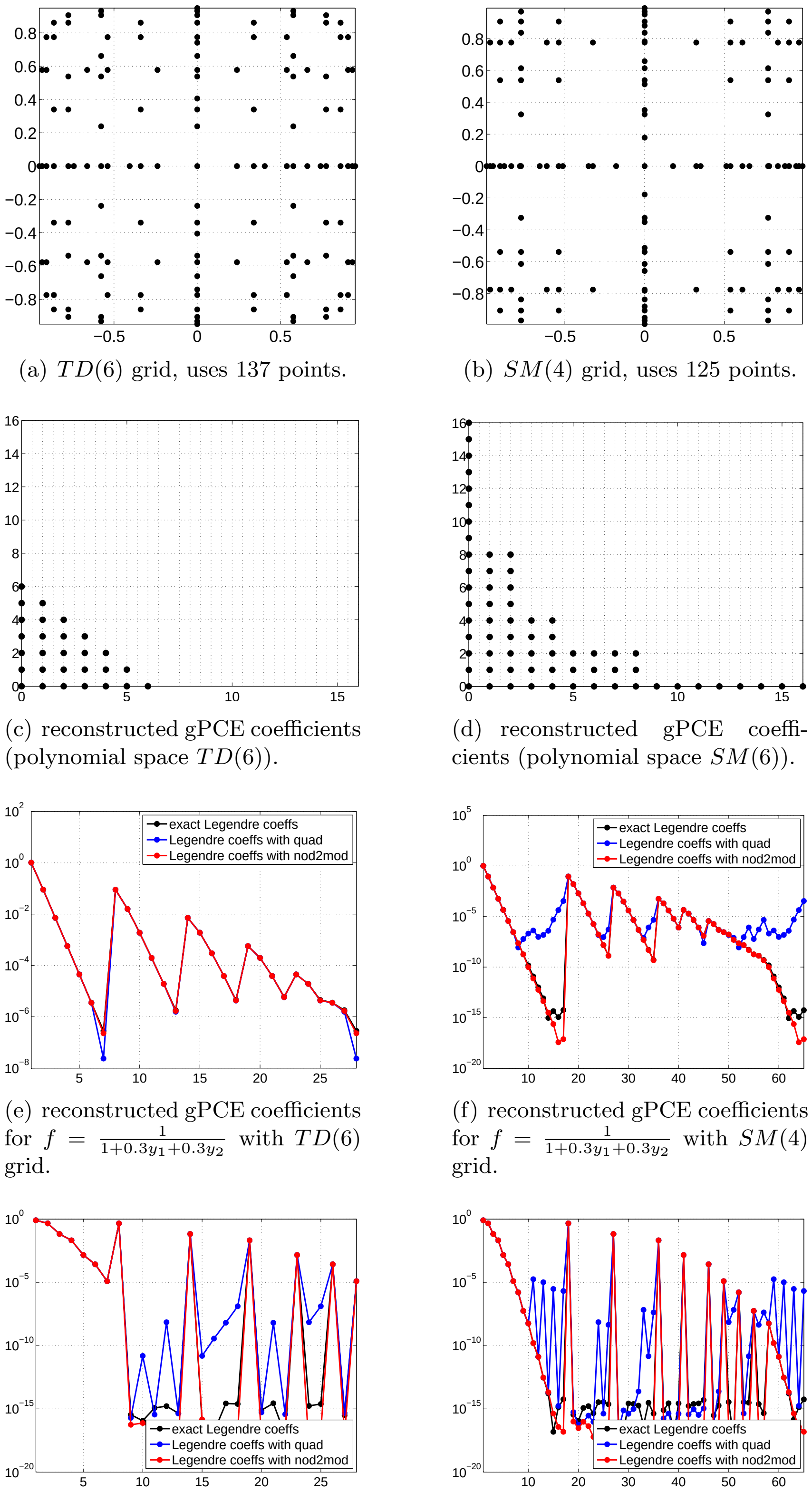
In general, the condition number of Q depends on $\mathcal{H}_n^{m(i_n)}$. However, the entries $\mathcal{L}_\mathbf{p}(\mathbf{y_j})$ can be precomputed, and the matrices Q can be assembled and possibly factorized once and for all. Finally, we mention that in the case of Clenshaw–Curtis abscissae the computation of $\boldsymbol{\beta}$ given the nodal values $\mathcal{L}_\mathbf{p}(\mathbf{y_j})$ can be performed efficiently using FFT techniques.

Figure 6.1 shows the results of the computation of the gPCE coefficients by converting a sparse grid approximation into a gPCE expansion, and by using the same sparse grid as a quadrature rule to approximate the integrals in (6.7). We compare the results obtained with two different sparse grids, both built using Gauss–Legendre quadrature points, namely $TD(6)$ and $SM(4)$ sparse grids. The reference values for the gPCE coefficients are computed with a very refined sparse grid quadrature rule. The results show that the computation of gPCE coefficients through the conversion approach is much more robust than the quadrature one, especially in the case of the SM grid. See [22] for additional numerical results.

6.3 The deterministic problem

In this Chapter we perform a global sensitivity analysis on a model describing the evolution over time of some relevant characteristics of a sedimentary basin, namely the porosity, the temperature and the pressure, whose knowledge is very relevant for the oil reservoir engineering practice.

In a nutshell, sedimentary basins form when sediments are deposited over long periods of time. Rocks form gradually as a result of the compaction of such sediments, which is caused both by mechanical stresses (the weight of the above layers of sediments), and chemical reactions, which are driven by temperature. One has also to take into account the explusion of the fluid from the sediments, which can be described as a Darcy flow driven by the pressure in the basin. The result of this combined action is a reduction of the porosity of the sediment, that is the free space among rock grains, which of course plays a crucial part in the determination of the rock permeability (the other factor being the disposition of the free space).



(a) $TD(6)$ grid, uses 137 points.

(b) $SM(4)$ grid, uses 125 points.

(c) reconstructed gPCE coefficients (polynomial space $TD(6)$).

(d) reconstructed gPCE coefficients (polynomial space $SM(6)$).

(e) reconstructed gPCE coefficients for $f = \frac{1}{1+0.3y_1+0.3y_2}$ with $TD(6)$ grid.

(f) reconstructed gPCE coefficients for $f = \frac{1}{1+0.3y_1+0.3y_2}$ with $SM(4)$ grid.

(g) reconstructed gPCE coefficients for $\sin(y_1 - \frac{1}{2}) + \sin(y_2 - \frac{1}{2})$ with $TD(6)$ grid.

(h) reconstructed gPCE coefficients for $\sin(y_1 - \frac{1}{2}) + \sin(y_2 - \frac{1}{2})$ with $SM(4)$ grid.

Figure 6.1: Comparison of the computation of the gPCE coefficients of a function via the modal conversion of a sparse grid approximation or sparse grid quadrature. Plots show the two grids used, the gPCE coefficients that can be reconstructed in both cases, and the numerical results for two different functions. The sparse grids are built using Gauss–Legendre quadrature points, and the gPCE coefficients are plotted in lexicographic order.

Among others, quartz precipitation in sandstones and smectite–illite transformation in shales are regarded as the predominant geochemical compaction processes ([12, 14, 67, 91]). In this work we will however focus on the study of quartz cementation in sandstones sediments only. Quartz cementation can be described as a sequence of three phases: dissolution of quartz grains, diffusion of the dissolved products and precipitation, where the last one is usually regarded as the rate limiting one, see e.g. [79]. Many conceptual models of quartz cementation at basin scale are available in literature; here we consider the model described in [105, 106], which proposes a simple empirical exponential law for quartz precipitation rate, and a minimal temperature (critical temperature) for the reaction to take place. This consitutive law has to be coupled with the Darcy law for the fluid explulsion, the temperature equation and the equation for the mechanical compaction of the sediments.

Most of the parameters of such model (physical properties of the liquid phase and solid porous medium, chemical parameters, geological information about the system) are affected by significant amounts of uncertainty, which is mainly related to the extreme difficulty to provide direct measurements of the quantities of interest at the space and time scales characterizing a basin compaction process.

6.3.a Mathematical formulation and discretization

We refer to the full paper [34] for a complete discussion on the details of the matemathical model and its discretization, which is beyonds the goals of this thesis. Here we only mention that we consider a monodimensional model along the vertical direction z (i.e. the depth of the basin). The complete model results in a coupled system of 7 equations, namely 3 conservation equations:

1. the mass conservation in divergence form;
2. the energy conservation, that is an elliptic equation for the temperature;
3. the force balance (algebraic balance);

and 4 constitutive equations, describing:

1. the permeability K as a function of the porosity ϕ, that can be tuned with two parameters k_1, k_2;
2. the fluid explusion flux as a function of the permeability (the well-known Darcy's law);
3. the thermal conductivity of the water/rock system;
4. the rate of change of the porosity, taking into account the mechanical and chemical processes,

$$\frac{d\phi}{dt} = \frac{d\phi_M}{dt} - \frac{d\phi_Q}{dt} \quad \phi > 0$$

   where ϕ_M is the rate of change for the porosity ascribable to mechanical compaction, that depends on a parameter β, and ϕ_Q is the rate of change of the porosity ascribable to the quartz precipitation chemical reaction, that is driven by two additional parameters a, b.

See table 6.1 for the full set of equations. As for the numerical discretization, we mention that the equations are solved in a Lagrangian framework, so that the grid moves along the z-axis and deforms following the thickness and position of the strata as time evolves, see figure 6.2. Therefore no solid mass is transferred among cells and there are no advection terms. New elements are added on the top as deposition occurs. The Darcy law and the fluid conservation equations are solved together with a mixed finite element method using $\mathbb{P}_0 - \mathbb{RT}_0$ finite elements, see [13], and the same type of discretization can be applied to a suitable reformulation of the temperature equation. Finally, we point out that at each time step the complete system is solved with an iterative fixed point method.

6.4 Sensitivity analysis

6.4.a Identifying the random parameters

We consider 7 parameters among all those appearing in the model formulation as uncertain, that is the parameter β of the mechanical compaction equation, the two parameters a and b driving the quartz cementation rate kinetics, the activation temperature of the geochemical process T_c, the depth of the sea over the basin h_{sea}, and the two parameters k_1, k_2 appearing in the relation betweeen porosity and permeability.

Since we don't have any a-priori knowledge on their probability distribution, we will assume each y_i to be uniformely distributed in a given uncertainty range $\Gamma_i = [l_i, u_i]$ and we denote with Γ

conservation equations	
$\dfrac{\partial \phi \rho_l}{\partial t} + \dfrac{\partial \left[\phi \rho_l u_l\right]}{\partial z} = 0$	water mass conservation law
$\dfrac{\partial \left(1-\phi\right) \rho_s}{\partial t} + \dfrac{\partial \left[\left(1-\phi\right) \rho_s u_s\right]}{\partial z} = q_{qrtz}$	solid mass conservation law • q_{qrtz} is the quartz production rate
$C_1 \dfrac{\partial T}{\partial t} + C_2 \dfrac{\partial T}{\partial z} - \dfrac{\partial}{\partial z} \left(K_T \dfrac{\partial T}{\partial z} \right) = q_T$	energy conservation law • $C_1 = \phi \rho_l c_l + (1-\phi) \rho_s c_s$ • $C_2 = \phi \rho_l c_l u_l + (1-\phi) \rho_s c_s u_s$ • q_T represents internal heat sources
$\sigma = -\displaystyle\int_{z_{top}}^{z} \left[\phi \rho_l + (1-\phi) \rho_s\right] g dz + s_0 - p_l$	force balance • g is the gravity acceleration • s_0 is the weight of the sea water column • σ is resulting effective stress on the solid matrix
constitutive equations	
$K = 10^{k_1 \phi - k_2 - 15}$	porosity/permeability law
$\phi(u_l - u_s) = -\dfrac{K}{\mu_l} \left(\dfrac{\partial p_l}{\partial z} - \rho_l g \right)$	Darcy law • μ_l is the water viscosity
$K_T(T) = \lambda_l^{\phi} [\lambda_s(T)]^{1-\phi}$	thermal conductivity of the water/rock system • $\lambda_s(T) = \lambda_0 / (1 + c_0 T)$
$\dfrac{d\phi}{dt} = \dfrac{d\phi_M}{dt} - \dfrac{d\phi_Q}{dt},\ \phi > 0$	porosity rate of change • $\dfrac{d\phi_M}{dt} = -\beta(\phi_0 - \phi_f) \exp(-\beta\sigma) \dfrac{d\sigma}{dt}$ • $\dfrac{d\phi_Q}{dt} = \dfrac{M_Q}{\rho_Q} A_0 \left(\dfrac{\phi}{\phi_{act}} \right) a 10^{bT}$

Table 6.1: Full set of equations for the geochemical compaction model. Here u denotes a velocity field, p a pressure field, and ρ a density. A subscript l denotes that such quantity refers to the water phase, while a subscript s to the solid phase.

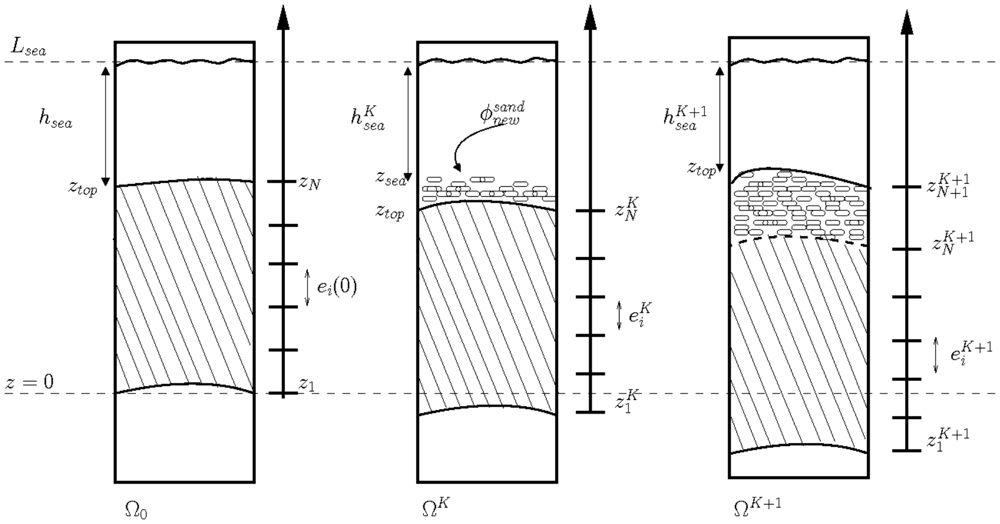


Figure 6.2: the computational grid moves according to the sedimental layers.

the hypercube $\Gamma = \Gamma_1 \times \Gamma_2 \ldots \times \Gamma_N$, so that every realization $\mathbf{y} \in \Gamma$. We will furthermore assume that all y_i are statistically independent; as a consequence, the joint probability density function of $\mathbf{y}$ over Γ is the product of uniform probability density functions in each direction,

$$\rho_\Gamma(\mathbf{y}) = \prod_{i=1}^{N} \rho_{\Gamma_i}(y_i) = \prod_{i=1}^{N} \frac{1}{u_i - l_i}. \tag{6.14}$$

All the other parameters introduced in the model are here considered fixed. Our choice is indeed arbitrary, as other involved parameters are manifestly affected by uncertainty; however, our aim here is to show the capability of our technique by including only the main parameters that affect the compaction process and their possible influence on pressure and temperature distribution. The tools provided may then be used to tackle a more detailed analysis, including a wider set of uncertain parameters.

Table 6.2 shows the uncertainty range associated to each uncertain parameter, and the literature source where each range has been assessed. The value of the sea depth h_{sea} (see Figure 6.2) is taken assuming a variation of about ±10% with respect to a reference value (500 m).

6.4.b Numerical results

We now analyze the results of the computation of the Sobol' indices, which have been obtained as explained in Section 6.2.

Figure 6.3(a) shows the distribution of porosity along the vertical axis at the final time-step, and two curves located at one standard deviation above and below the mean. As expected, it cleary appears that the deeper strata are characterized by a larger uncertainty. Note that the curve is not significative in the deepest strata ($z < -4000$), where the porosity approaches zero.

Figure 6.3(b) shows the global Sobol'indices (6.6) for each of the uncertain parameters along the vertical direction, where three regions can be distinguished. In the upper zone the porosity is only influenced by the boundary datum h_{sea} and by the mechanical part of the compaction process. Around a burial depth $z = -2000$m the quartz precipitation starts, so the rate of porosity reduction with burial depth increases, and the contribution of the chemical compactation is significative up to $z = -4000$. Note that the activation temperature only plays a role in the earliest strata of

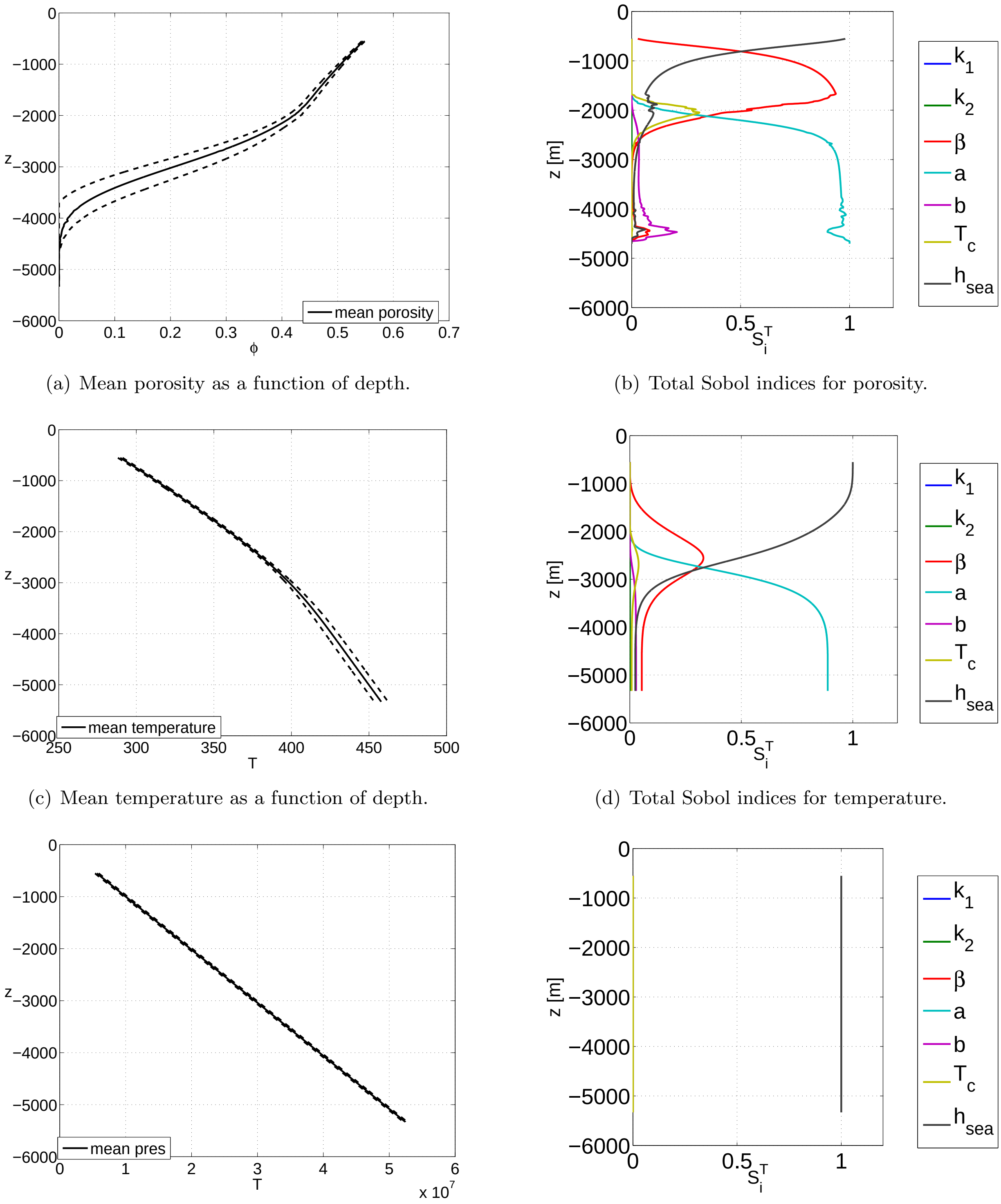
Parameter	l_i	u_i	**Bibliographic ref.**
$\beta[\text{Pa}^{-1}]$	5×10^{-8}	7×10^{-8}	[59]
$a[\text{mol m}^{-2}\text{ s}^{-1}]$	0.5×10^{-18}	3.5×10^{-18}	[105]
b $[\text{C}^{-1}]$	0.020	0.024	[105]
T_c [C]	70	90	[59]
h_{sea}[m]	450	550	–
k_1 [-]	14.07	14.22	[109]
k_2 [-]	1.35	2.38	[109]

Table 6.2: list of uncertain parameters. Each one is modeled as a uniform random variable ranging between l_i and u_i.

this second part, that is $-2000 < z < -1800$. In the deepest layers of the basin ($z < -4000$) all the pore space is filled with quartz due to geochemical compaction and the porosity approaches zero. In this region the Sobol indices are not significant due to the negligible values assumed by the porosity itself and exhibit an oscillatory behavior. This feature of the solution is induced by the fact that we consider here a homogeneous material where quartz precipitation is possible everywhere and we neglect possible grain coating effects. The computational model proposed here thus seems to be quite effective in reproducing the experimental observations. The results also suggest that in principle it would be possible to save some computational time by switching off the equations describing the chemical compaction in the upper layers of the computational domain, and moreover indicate that only a subset of the selected parameters has a significant influence of the system.

Figure 6.3(c)-6.3(d) shows the vertical distribution of mean temperature and the related indices. As expected, the temperature increases with the depth, and again the h_{sea} parameter, the mechanical and geochemical compaction processes influence the state variable, the former being the only important effect in the most superficial layers ($z > -2000$). Results also suggest that all the variability of the temperature is induced by different thermal diffusivity associated with the solid and the liquid phases, hence it is intrinsically linked with the porosity field.

The vertical distribution of pressure and the Sobol indices are shown in Figure 6.3(e)-6.3(f). The pressure distribution appears to be basically linear with depth. All uncertainty in fluid pressure distribution is associated with the h_{sea} parameter and no feedback of the geochemical compaction is observed. This is likely due to the homogeneous material assumption.



(a) Mean porosity as a function of depth.

(b) Total Sobol indices for porosity.

(c) Mean temperature as a function of depth.

(d) Total Sobol indices for temperature.

(e) Mean pressure as a function of depth.

(f) Total Sobol indices for pressure.

Figure 6.3: Mean values for porosity, temperature and pressure at final time (left column), with dotted lines one standard deviation above and below, and corresponding Sobol'indices (right column).

Conclusions and perspectives

In this thesis we have investigated the theoretical aspects of the polynomial approximation of the solution u of a PDE with stochastic parameters. To this end, one has to consider a parametrization of the underlying probability space with N random variables. As a consequence, the PDE is rewritten as a parametric problem depending on N parameters and it is possible to compute a polynomial surrogate model for u.

We have considered two different means of obtaining such surrogate model: a projection one (the Stochastic Galerkin Method), and a collocation one (the Stochastic Collocation Method), which we have compared in detail in Chapter 2. The most relevant feature of our work in this sense is the equivalence theorem that allows us to set the Stochastic Galerkin and Collocation methods in the same polynomial space, which is actually a crucial aspect if one wants to obtain a fair comparison between the two techniques.

We have also proposed some new algorithms in the context of the Galerkin and Collocation techniques that allow in some situations to improve significantly the error decay, see Chapters 3-4 and 5. It is important to remark that in particular the "Optimal Sets" techniques (Chapters 3 and 4) are capable of working formally with a countable infinite number of parameters, as the less influent ones will be automatically discarded in an adaptive way.

The analysis of such methods is far from being complete. The analysis of a-priori error estimates for the "Optimal sets" approach should be tackled, both for the Galerkin and the Collocation approaches. Moreover, it will be interesting to improve the estimate of the error contribution of the hierarchical surpluses, that we have addressed in this thesis only with a numerical/heuristical approach. A future work to analyze is also to extend the work on the optimal sparse grids to the case of non-nested interpolation points. A convergence analysis should also be addressed for the *PGD* method presented in Chapter 5.

Yet the results obtained in practical applications (Chapters 4, 6) are encouraging, although the cases considered here are still simplified. It will be interesting to extend the results on the groundwater flows to situations where the covariance of the log-permeability is modeled as as exponential covariance rather than gaussian, and the correlation length is smaller. In such situations we expect that the solution may feature low regularity with respect to the random parameters (due to the low regularity of the exponential covariance function), and thus ad-hoc enhancements of the sparse grid methods should be devised. This would lead the way for interesting engineering applications, like the assessment of the catchement area of a well and the pollutant remediation. Note that in realistic applications the number of random variables to be considered in the model could be higher than the cases considered in this work, which represents a challenge also from an algorithmic and computational point of view.

Finally, we remark that in this thesis we have mostly considered elliptic problems. More general situations have been addressed in Chapters 5 and 6, and it would be of interest to extend to such situations the optimal sets procedure developed in Chapter 3. To the same end the sparse-grid-based sensitivity analysis techniques developed in Chapter 6 should also be further analyzed.

Appendix A

Addendum on the convergence of optimal sets technique

February 5, 2012

In this addendum we provide some convergence results for the optimal sets technique, both in the Stochastic Collocation and Galerkin case. For Stochastic Collocation we provide a characterization of the error for the "optimal" sparse grid in terms of weighted ℓ^p summability of the profits. This result extends to the case of sparse grids the known result on non-linear approximation (see e.g. [20, 21] and references therein). For Stochastic Galerkin, we analyze the convergence in the specific case of the "inclusions" test described in Chapter 2 and reinterpret the numerical results there obtained in view of the estimates shown here.

We will work in the same setting as Chapters 2-3. Thus, we consider a set of N independent random variables y_i, uniformly distributed over $\Gamma_i = [-1, 1]$, with joint probability density $\rho(\mathbf{y}) = 1/2^N$, and we denote the stochastic domain as $\Gamma = \Gamma_1 \times \Gamma_2, \ldots, \times\Gamma_N$. Next, we denote as $u \in H_0^1(D) \otimes L^2_\rho(\Gamma)$ the weak solution of

$$\begin{cases} -\operatorname{div}(a(\mathbf{x},\mathbf{y})\nabla u(\mathbf{x},\mathbf{y})) = f(\mathbf{x}) & \mathbf{x} \in D \\ u(\mathbf{x},\mathbf{y}) = 0 & \mathbf{x} \in \partial D \end{cases} \qquad \rho(\mathbf{y})d\mathbf{y} \text{ a.e. in } \Gamma, \tag{A.1}$$

where the diffusion coefficient a is such that there exist two constants a_{min}, a_{max} such that

$$0 < a_{min} < a(\mathbf{x},\mathbf{y}) < a_{max} < \infty$$

for almost every $\mathbf{x} \in D$ and $\mathbf{y} \in \Gamma$. As in the previous Chapters, we will exploit the fact that u can be understood either as a function in the tensor space $H_0^1(D) \otimes L^2_\rho(\Gamma)$ or as a $H_0^1(D)$-valued square-integrable function of $\mathbf{y} \in \Gamma$, i.e. $u \in L^2_\rho(\Gamma; H_0^1(D))$, and use the best notation depending on the situation. Finally, we recall that u can be expanded in a Legendre series, that reads

$$u(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{p}\in\mathbb{N}^N} u_{\mathbf{p}}(\mathbf{x})\mathcal{L}_{\mathbf{p}}(\mathbf{y}), \tag{A.2}$$

where $\mathcal{L}_{\mathbf{p}}$ are the N-variate Legendre polynomials, $\mathcal{L}_{\mathbf{p}}(\mathbf{y}) = \prod_{n=1}^N L_{p_n}(y_n)$, and $L_{p_n}(y_n)$ are the monovariate orthonormal Legendre polynomials in y_n. We also need to introduce the standard $L^\infty(\Gamma)$-normalized Legendre polynomials $\mathscr{L}_j$, for which the following properties hold:

- $\mathscr{L}_j(1) = 1$;
- $\int_{-1}^1 \mathscr{L}_j(t)\mathscr{L}_k(t)dt = \delta_{jk}(j + 1/2)^{-1}$;
- $L_j(t) = \sqrt{2j+1}\mathscr{L}_j(t)$.

A.1 Convergence of the quasi-optimal sets Galerkin method

As discussed in Chapter 3, the optimal M dimensional polynomial space for the Stochastic Galerkin method is the one spanning the Legendre polynomials corresponding to the M largest coefficients in the Legendre expansion (A.2). This choice indeed minimizes the projection error

$$\|u - \sum_{\mathbf{p} \in \mathcal{S}_M} u_{\mathbf{p}} \mathcal{L}_{\mathbf{p}} \|^2_{V \otimes L^2_\rho(\Gamma)} = \sum_{\mathbf{p} \notin \mathcal{S}_M} \|u_{\mathbf{p}}\|^2_V. \tag{A.3}$$

A possible strategy to assess the convergence of the resulting generalized Polynomial Chaos Expansion of u is to order the Legendre coefficients $\|u_{\mathbf{p}}\|^2_V$ in decreasing order according to a suitable a-priori estimate and study the summability properties of the sequence thus obtained. This idea has been investigated e.g. in [20, 21] for the case when the diffusion coefficient can be written as $a(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^\infty y_i b_i(\mathbf{x})$, with y_i uniform random variables over $[-1, 1]$ and $\{\|b_i\|_\infty\}_{i \in \mathbb{N}} \in \ell^p$ for some $p < 1$.

1.1.a Direct estimates for a special case of interest

In this Section we will follow a different approach, and restrict our focus to the case in which u obeys the following Assumption:

Assumption 1. *The complex continuation of u, $u^* : \mathbb{C}^N \to H^1_0(D)$ is a $H^1_0(D)$-valued holomorphic function in the polydisc*

$$E_S = \bigotimes_{n=1}^N E_{n,S_n}, \quad E_{n,S_n} = \{z_n \in \mathbb{C} : |z_n| < S_n\}$$

with $\sup_{\mathbf{z} \in E_S} \|u^*(\mathbf{z})\|_{H^1_0(D)} \leq B_u$.

We will see that this class of functions includes e.g. the solution of the inclusions test investigated in Chapter 2. We start by proving a result on the decay of the coefficients of the Legendre expansion (A.2) for u satisfying Assumption 1. To this end, we first need the following simple Lemma.

Lemma 2. *The polyellipse $\mathcal{E}_S = \bigotimes_{n=1}^N \mathcal{E}_{n,S_n}$, with*

$$\mathcal{E}_{n,S_n} = \left\{ z_n \in \mathbb{C} : \mathfrak{Re}\,(z) = \frac{\varrho_n + \varrho_n^{-1}}{2} \cos\phi,\ \mathfrak{Im}\,(z) = \frac{\varrho_n - \varrho_n^{-1}}{2} \sin\phi,\ \phi \in [0, 2\pi) \right\}$$

and $\varrho_n = S_n + \sqrt{S_n^2 - 1}$ is included in the polydisc E_S.

Proof. The value of ϱ_n is obtained enforcing $z_n = S_n$ to belong to $\mathcal{E}_{n,S_n}$, i.e. $\frac{\varrho_n + \varrho_n^{-1}}{2} = S_n$. In this way the semi-major axis of E_{n,S_n} and $\mathcal{E}_{n,S_n}$ coincide. The inclusion $\mathcal{E}_S \subset E_S$ holds since one can easily verify that the length of the semi-minor axis of each $\mathcal{E}_{n,S_n}$ is smaller that S_n. □

We are now in position to prove the following estimate on the Legendre coefficients.

Proposition 3. *If u fulfills Assumption 1, the coefficients of the Legendre expansion* (A.2) *decay as*

$$\|u_{\mathbf{p}}\|_{H^1_0(D)} \leq C_{Leg}\, e^{-\sum_{n=1}^N g_n p_n} \prod_{n=1}^N \sqrt{2p_n + 1}, \tag{A.4}$$

with $g_n = \log(\varrho_n)$ and $C_{Leg} = B_u \prod_{n=1}^N \frac{L(\mathcal{E}_{n,S_n})}{4(\varrho_n - 1)}$.

Here $L(\mathcal{E}_{n,S_n})$ denotes the length of the ellipse $\mathcal{E}_{n,S_n}$ in Lemma 2, ϱ_n is as in Lemma 2, and B_u is as in Assumption 1.

Proof. The proof follows closely the argument in [25, Section 12.4]. From Assmuption 1 and Lemma 2 we have that u is analytic in the region inside $\mathcal{E}_S$, and hence we can exploit the Cauchy's formula to rewrite the $\mathbf{p}$-th Legendre coefficient as

$$u_{\mathbf{p}} = \int_\Gamma u(\mathbf{x},\mathbf{y})\mathcal{L}_{\mathbf{p}}(\mathbf{y})\rho(\mathbf{y})d\mathbf{y} = \int_\Gamma \mathcal{L}_{\mathbf{p}}(\mathbf{y})\rho(\mathbf{y}) \oint_{\mathcal{E}_S} \frac{u^*(\mathbf{x},\mathbf{z})}{\prod_n 2\pi i(z_n - y_n)} d\mathbf{z}d\mathbf{y}$$
$$= \oint_{\mathcal{E}_S} u^*(\mathbf{x},\mathbf{z}) \prod_{n=1}^N \frac{1}{2} \int_{\Gamma_n} \frac{L_{p_n}(y_n)}{2\pi i(z_n - y_n)} dy_n d\mathbf{z}.$$

Next, let

$$Q_{p_n}(z_n) = \int_{\Gamma_n} \frac{\mathscr{L}_{p_n}(y_n)}{(z_n - y_n)} dy_n\,.$$

From [25, Lemma 12.4.6] it follows that for all $z_n \in \mathcal{E}_{n,S_n}$,

$$|Q_{p_n}(z_n)| \leq \pi \frac{(1/\varrho_n)^{p_n}}{\varrho_n - 1}.$$

Then it holds

$$\|u_{\mathbf{p}}\|_{H_0^1(D)} \leq \sup_{\mathcal{E}_S} \|u^*\|_{H_0^1(D)} \prod_{n=1}^N \frac{\sqrt{2n+1}}{4\pi} \oint_{\mathcal{E}_{n,S_n}} |Q_{p_n}(z_n)| dz_n$$
$$\leq \sup_{\mathcal{E}_S} \|u^*\|_{H_0^1(D)} \prod_{n=1}^N \frac{\sqrt{2n+1}}{4\pi} \pi \frac{(1/\varrho_n)^{p_n}}{\varrho_n - 1} \oint_{\mathcal{E}_{n,S_n}} dz_n$$
$$\leq \sup_{\mathcal{E}_S} \|u^*\|_{H_0^1(D)} \prod_{n=1}^N \frac{\sqrt{2n+1}\, L(\mathcal{E}_{n,S_n})}{4(\varrho_n - 1)} e^{-p_n \log(\varrho_n)}.$$

Finally observe that, since u satisfies Assumption 1 and $\mathcal{E}_S \subset E_S$,

$$\sup_{\mathcal{E}_S} \|u^*\|_{H_0^1(D)} \leq \sup_{E_S} \|u^*\|_{H_0^1(D)} \leq B_u.$$

□

Remark 4. *The square root factor in* (A.4) *is of course negligible compared to the exponential decreasing term* $e^{-\sum_n g_n p_n}$. *We can therefore assume that the simplified expression*

$$\|u_{\mathbf{p}}\|_{H_0^1(D)} \leq \prod_{n=1}^N \widehat{C}_{Leg} e^{-\widehat{g}_n p_n} \tag{A.5}$$

is also an accurate estimate for the Legendre coefficients of u *satisfying Assumption 1, at the price of substituting* g_n *with* $\widehat{g}_n < g_n$ *and, possibly,* C_{Leg} *with* $\widehat{C}_{Leg} > C_{Leg}$. *In particular, it holds* $\sqrt{2p+1} \leq C(\epsilon, g)e^{\epsilon g p}$ *for every* $\epsilon > 0$, *with* $C(\epsilon, g) \to +\infty$ *as* $\epsilon \to 0$.

From Chapter 3, we know that if the Legendre coefficients of u decay as in equation (A.5), the family of (anisotropic) TD sets, $TD(w) = \{\mathbf{p} \in \mathbb{N}^N : \sum_{n=1}^N g_n p_n \leq w\}$, is a sharp estimate of the optimal polynomial space for the Galerkin method, that we denote as $\mathbb{P}_{TD(w)}(\Gamma)$ (see Section 3.3.a for details). In the following $u_w = \sum_{\mathbf{p} \in TD(w)} u_{\mathbf{p}} \mathcal{L}_{\mathbf{p}} \in H_0^1(D) \otimes \mathbb{P}_{TD(w)}(\Gamma)$ will denote the TD Galerkin approximation of u.

Following closely the argument in [71], we can prove the convergence estimate for the TD approximation of u. Such argument is for isotropic problems only, so we introduce a further assumption on u.

Assumption 5. *Assumption 1 holds with $S_n = S$, for $n = 1, \dots, N$.*

As a consequence, the parameters ϱ_n describing the polyellipse in Lemma 2 are all equal, as well as the coefficients g_n driving the decay of the Legendre coefficients in Proposition 3/Remark 4, and thus the optimal polynomial space is indeed the isotropic TD, $TD(w) = \{\mathbf{p} \in \mathbb{N}^N : \sum_{n=1}^N p_n \leq w\}$.

The first step to prove the convergence of the TD approximation is to state the following theorem, which gives the optimality of the Galerkin approximation.

Theorem 6. *It holds*

$$\|u - u_w\|_{L^2_\rho(\Gamma;H^1_0(D))} \leq C_{opt} \inf_{v \in H^1_0(D) \otimes \mathbb{P}_{TD(w)}(\Gamma)} \|u - v\|_{L^2_\rho(\Gamma;H^1_0(D))},$$

where C_{opt} is a constant depending on a_{min}, a_{max}.

Proof. The proof is an immediate rewriting of [71, Theorem 1]. □

Note that indeed such Theorem does not require Assumption 5. Next, we shall need the following Lemma (see [6] for a proof), which conversely relies on Assumption 5.

Lemma 7. *Suppose u satisfies Assumptions 1 - 5, and let $\mathcal{M}_{u,w}$ be the Maclaurin polynomial of u on the complex domain,*

$$\mathcal{M}_{u,w}(\mathbf{z}) = \sum_{\mathbf{p} \in TD(w)} \alpha_{\mathbf{p}} \prod_{n=1}^N z_n^{p_n}.$$

Then, for $0 < R < S$,

$$\sup_{\mathbf{z} \in E_R} \|u^*(\mathbf{z}) - \mathcal{M}_{u,w}(\mathbf{z})\|_{H^1_0(D)} \leq \frac{B_u}{S/R - 1} e^{-\widetilde{g} w},$$

with B_u as in Assumption 1 and $\widetilde{g} = \log \frac{S}{R}$.

The convengence rate for the isotropic TD approximation can then be estimated combining Theorem 6 and Lemma 7.

Theorem 8. *Suppose that u satisfies Assumptions 1 and 5, with $\Gamma \subset E_S$ for some $S > 1$.Then it holds*

$$\|u - u_w\|_{L^2_\rho(\Gamma;H^1_0(D))} \leq C_{opt} \frac{B_u}{S - 1} e^{-\widetilde{g} w}, \tag{A.6}$$

with B_u as in Lemma 7, $\widetilde{g} = \log S$, and C_{opt} as in Theorem 6.

Proof. We use Lemma 7 with $R = 1$ (note that the intersection of E_1 with the real axis is Γ). Then we have

$$\begin{aligned}\|u - u_w\|_{L^2_\rho(\Gamma;H^1_0(D))} &\leq C_{opt} \inf_{v \in H^1_0(D) \otimes \mathbb{P}_{TD(w)}(\Gamma)} \|u - v\|_{L^2_\rho(H^1_0(D);\Gamma)} \\ &\leq C_{opt} \|u - \mathcal{M}_{w,u}\|_{L^2_\rho(H^1_0(D);\Gamma)} \\ &\leq C_{opt} \|u - \mathcal{M}_{w,u}\|_{L^\infty(H^1_0(D);\Gamma)} \leq C_{opt} \frac{B_u}{S - 1} e^{-\widetilde{g} w}.\end{aligned}$$

□

Theorem 8 states an exponential convergence of the error with respect to the degree of the polynomial approximation. In practice however one is more concerned with the convergence of u_w

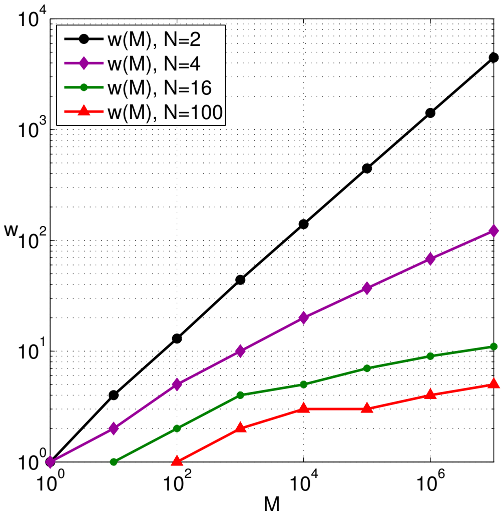


Figure A.1: $w(M)$ for different values of N.

with respect to the number of degrees of freedom, i.e. to the number M of polynomials in $TD(w)$. Hence, we are lead to the problem of finding a bound for the function $w = w(M)$. Note that the inverse of such function, $M = M(w)$, is known analytically, $M = \binom{N+w}{N}$. The function $w(M)$ can thus be easily computed numerically: it is of course increasing in M and decreasing in N, i.e. the level w needed to have M terms in the set is lower for higher N, see Figure A.1. In general, using the quite crude inequality $M \leq e^{w(1+\log N)}$, which has been shown in [71, eq. 25], one can obtain a bound for $e^{-\widetilde{g}w}$, though not sharp, as $e^{-\widetilde{g}w} \leq M^{-\widetilde{g}/(1+\log N)}$.

1.1.b The inclusions problem

We now consider again the inclusions problem examined in Chapter 2. We recall that for such problem the diffusion coefficient in (A.1) is given by

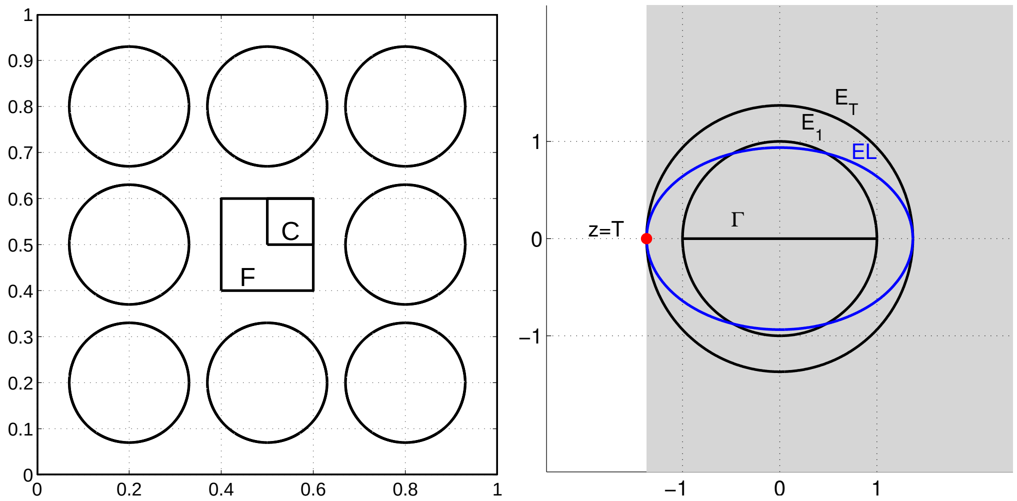
$$a(\mathbf{x}, \mathbf{y}) = 1 + \sum_{n=1}^{N} \chi_n(\mathbf{x}) y_n, \tag{A.7}$$

where $\chi_n(\mathbf{x})$ are the indicator functions of the disjoint circular subdomains $D_n \subset D = [0,1]^2$ as in Figure A.2(a), and y_n are independent random variables uniformely distributed in $[y_{min}, y_{max}]$. We will first prove that the TD sets are quasi-optimal sets for such problem (i.e. we can apply Propostion 3). Then, since the problem is trivially isotropic and hence satisfies Assumption 5, we will apply Theorem 8 and show that the numerical results obtained for such problem are in agreement with the predicted convergence rate.

We shall begin by reparametrizing the diffusion coefficient in terms of new random variables distributed over $[-1, 1]$, so that we can apply the discussion of the previous Section. For the sake of notation, we will still denote the new variables as y_i, i.e. $y_i \sim \mathcal{U}(-1, 1)$. The new diffusion coefficient will be therefore

$$a(\mathbf{x}, \mathbf{y}) = 1 + \sum_{n=1}^{N} \chi_n(\mathbf{x}) \frac{y_n + 1}{2}(y_{max} - y_{min}) + y_{min}. \tag{A.8}$$

We can now prove the following lemma on the analyticity region of u, that we denote by Σ.



(a) Physical domain for the inclusions problem.

(b) Analyticity regions along the n-th direction. The gray area denotes the analyticity region Σ. $z = T$ is the singularity up to which it is possible to extend u^*. E_1 and E_T are the circles used to prove the convergence of TD estimates (Proposition 8), while EL is the ellipse used to estimate the decay of the Legendre coefficients (Proposition 3).

Figure A.2: Inclusions problem.

Lemma 9. *The solution u to* (A.1) *corresponding to a diffusion coefficient* (A.8) *is analytic in the region*

$$\Sigma = \bigotimes_{n=1}^{N} \Sigma_n, \quad \Sigma_n = \{z_n \in \mathbb{C} : \mathfrak{Re}\,(z_n) > T\}\,, \tag{A.9}$$

with $T = \dfrac{2 + y_{max} + y_{min}}{y_{min} - y_{max}}$.

<u>Proof.</u> It has been pointed out in Chapters 2, 3 that the solution of an elliptic problem as (A.1) is analytic in each direction y_n. In particular, having fixed the values of all the random variables but the n-th, let us write $a_n^*(\mathbf{x}, z_n) = a(\mathbf{x}, y_1, y_2, \dots, y_{n-1}, z_n, y_{n+1}, \dots, y_N)$ and $u_n^*(\mathbf{x}, z_n) = u(\mathbf{x}, y_1, y_2, \dots, y_{n-1}, z_n, y_{n+1}, \dots, y_N)$. Such u_n can be extended in $\Sigma_n = \{z_n \in \mathbb{C} : \mathfrak{Re}\,(z_n) > T\}$ where T is computed as the value such that

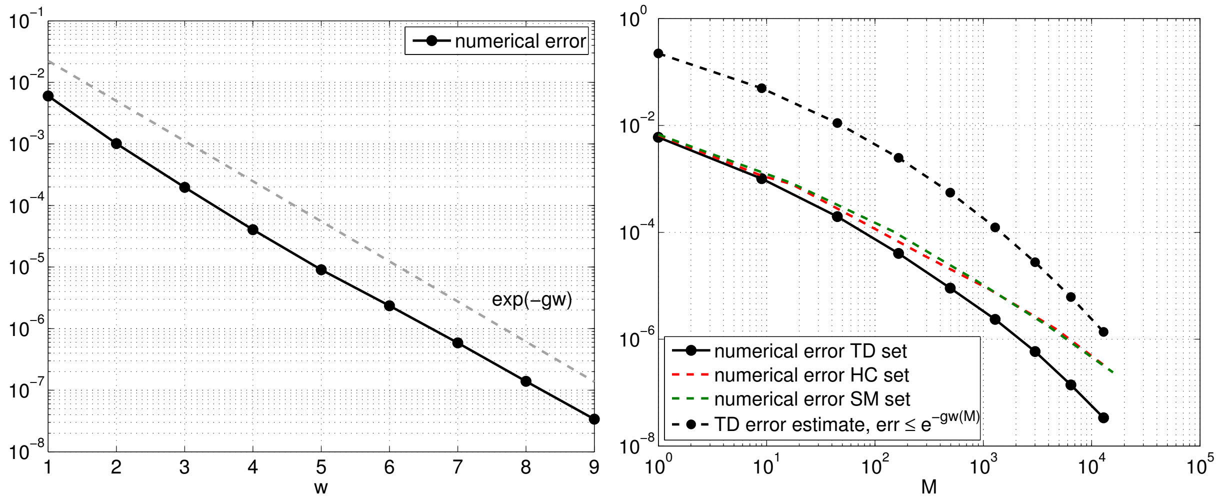
$$\exists\, \mathbf{x} \in D : a_n(\mathbf{x}, T) = a(\mathbf{x}, y_1, y_2, \dots, y_{n-1}, T, y_{n+1}, \dots, y_N) = 0.$$

This amounts to

$$1 + \frac{T+1}{2}(y_{max} - y_{min}) + y_{min} = 0,$$

whose solution is $T = (2 + y_{max} + y_{min})/(y_{min} - y_{max})$. Note that since the subdomains D_n do not overlap, $a_n(\mathbf{x}, T) = 0$ in D_n only, i.e. T does not depend on the value of any of the other random variable y_i. Thus, the analyticity region of u is the tensor product of the analyticity regions Σ_n. □

In particular, it is straigthforward to build a polydisc E_S and a polyellipse $\mathcal{E}_S$ inside Σ_n, (see Figure A.2(b)). Thus, the discussion of the previous Section can be applied to the inclusion problem, and we have the following final result:



(a) Error convergence with respect to w.

(b) Error convergence with respect to M. The values $w(M)$ are computed numerically.

Figure A.3: Comparison between estimate (A.6) and the numerical results from Chapter 2. The rate $g = 1.5$ has been numerically assessed.

Theorem 10. *The Legendre coefficients of the solution of the inclusions problem decay as*

$$u_{\mathbf{p}} \leq C(\epsilon) e^{-(1-\epsilon) g \sum_{n=1}^{N} p_n}, \quad \forall\, 0 < \epsilon < 1,$$

with $g = \log(\mathcal{T} + \sqrt{\mathcal{T}^2 - 1})$, *and* $\mathcal{T} = |T|$. *Hence, the polynomial space* $\mathbb{P}_{TD(w)}(\Gamma)$ *is the quasi-optimal space for the Galerkin method, and it holds*

$$\|u - u_w\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq C_{opt} \frac{B_u}{\mathcal{T} - 1} e^{-\widetilde{g} w}, \tag{A.10}$$

with B_u *as in Assumption 1 and* $\widetilde{g} = \log \mathcal{T}$*.*

Proof. The analyticity region (A.9) for u includes the polydisc $E_{\mathcal{T}}$, hence u satisfies Assumption 1 and we can apply Proposition 3/Remark 4 to estimate the decay of the coefficients $u_{\mathbf{p}}$. The optimality of $\mathbb{P}_{TD(w)}(\Gamma)$ then derives from the discussion in Section 3.3.a, and we can apply Theorem 8 with $S = \mathcal{T}$ to estimate its convergence rate. □

We now reconsider the numerical results obtained in Chapter 2 in view of the Theorem just proved. Figure A.3(a) shows the convergence with respect to w of the $L^2_\rho(\Gamma)$ error for the TD approximation of $\psi_1(u)$, with $\psi_1 : H^1_0(D) \to \mathbb{R}$ linear functional of u (see Section 2.4.a for details on the definition of ψ_1), and shows an optimal correspondence between the numerical results and the theoretical estimate (A.10) in Theorem 10. Note however that the rate g observed experimentally is $\widetilde{g} \approx 1.5$, which is much higher than the theoretically predicted, which amounts to $\widetilde{g} = \log \mathcal{T} \approx 0.025$. Thus the estimate we provide captures the right behaviour of the error convergence (i.e. exponential), but is very conservative. Yet, it can still provide the ansatz for a tuned estimate, which is what we advertise.

Figure A.3(b) shows instead the convergence with respect to M for different polynomial approximations, namely TD, HC and SM (see Section 2.3). As already observed in Chapter 2, the TD approximation is the most efficient approximation scheme for the problem of interest, and now can be also understood as the optimal approximation.

A.2 A convergence estimate for the quasi-optimal sparse grids for Stochastic Collocation

In this section we derive an a-priori estimate for the optimal set sparse grid. To this end, we start by briefly recalling the notation needed. We consider a sequence of index sets $\mathcal{I}(w) \in \mathbb{N}_+^N$ with the following properties:

- $\mathcal{I}(0) = \{(1, 1, \dots, 1)\}$,
- $\mathcal{I}(w) \subset \mathcal{I}(w+1)$,
- $\forall\, \mathbf{i} \in \mathcal{I}, \quad \mathbf{i} - \mathbf{e}_j \in \mathcal{I}$ for $1 \leq j \leq N, i_j > 1$ (admissibility condition, see e.g. Section 3.4).

The sparse grid approximation of $u \in H_0^1(D) \otimes L_\rho^2(\Gamma)$ is defined as

$$u_w = \mathcal{S}_{\mathcal{I}(w)}^m[u] = \sum_{\mathbf{i} \in \mathcal{I}(w)} \bigotimes_{n=1}^N \Delta_n^{m(i_n)}[u], \quad \Delta_n^{m(i)}[u] = \mathcal{U}_n^{m(i)}[u] - \mathcal{U}_n^{m(i-1)}[u], \tag{A.11}$$

where $m(0) = 0$, $m(1) = 1$, $m(i) < m(i+1)$, and $\mathcal{U}_n^{m(i_n)} : L_{\rho_n}^2(\Gamma_n) \to \mathcal{C}^0(\Gamma_n)$ is an interpolant operator over $m(i_n)$ points along the n-th direction of the stochastic domain. Let us also denote with $W_{\mathcal{I}(w),m}$ the total work of the sparse grid (A.11), i.e. the total number of interpolation points used by (A.11).

As shown in Chapter 3, the optimal grids are built using nested points. We will focus here to the choice of Clenshaw–Curtis interpolation points and their corresponding counting function m defined as

$$m(i) = \begin{cases} 0 \text{ if } i = 0 \\ 1 \text{ if } i = 1 \\ 2^{i-1} + 1, \text{ if } i > 1, \end{cases} \tag{A.12}$$

see Chapter 3 for details. Following [72], it is useful to introduce the function $r(i) = m(i) - m(i-1)$, i.e.

$$r(i) = \begin{cases} 1 \text{ if } i = 1 \\ 2 \text{ if } i = 2 \\ 2^{i-2} \text{ if } i > 2. \end{cases} \tag{A.13}$$

Next, for each multiindex $\mathbf{i} \in \mathcal{I}(w)$ we introduce the operator $\Delta^{m(\mathbf{i})} = \bigotimes_{n=1}^N \Delta^{m(i_n)}$ (hierarchical surplus), and we associate to each of these operators an error contribution, a work contribution and a profit as follows:

- $\Delta E(\mathbf{i}) = \left\| \mathcal{S}_{\{\mathcal{J} \cup \mathbf{i}\}}^m[u] - \mathcal{S}_{\mathcal{J}}^m[u] \right\|_{V \otimes L_\rho^2(\Gamma)} = \left\| \Delta^{m(\mathbf{i})}[u] \right\|_{V \otimes L_\rho^2(\Gamma)}$, (A.14)
- $\Delta W(\mathbf{i}) = W_{\{\mathcal{J} \cup \mathbf{i}\},m} - W_{\mathcal{J},m}$, (A.15)
- $P(\mathbf{i}) = \dfrac{\Delta E(\mathbf{i})}{\Delta W(\mathbf{i})}$, (A.16)

where $\mathcal{J}$ is any set of indices such that $\mathbf{i} \notin \mathcal{J}$ and $\{\mathcal{J} \cup \mathbf{i}\}$ is admissible (see Section 3.4.a for details).

The first step for the construction of the optimal sparse grid is to order the profits in decreasing order. For sake of notation we will still denote this sequence as $\{P_j\}_{j \in \mathbb{N}_+}$, with

$$P_j \geq P_{j+1}. \tag{A.17}$$

Note that this sequence may not be unique: in this case, any criterion to select a specific sequence satisfying (A.17) can be used.

It is useful to introduce a function that assigns to the j-th profit the corresponding multiindex and its inverse. With a slight abuse of notation, we will denote the former as $\mathbf{i}(j)$, and the latter as $j(\mathbf{i})$, i.e. $P_{\mathbf{i}(j)} \geq P_{\mathbf{i}(j+1)}$. We will also need to construct a sequence of error contributions and of work contributions, using the same order as the sequence of profits. We thus obtain the sequences $\{\Delta E_j\}_{j \in \mathbb{N}_+}$ and $\{\Delta W_j\}_{j \in \mathbb{N}_+}$: note that these sequences will not be ordered in general.

The optimal sparse grid is then built over the family of sets that collect the w most profitable indices,

$$\mathcal{I}(w) = \{\mathbf{i}(1), \mathbf{i}(2), \ldots, \mathbf{i}(w)\}. \tag{A.18}$$

We now make the following restrictive assumption:

Assumption 11. *Each of the sets in the sequence $\mathcal{I}(w)$ is admissible.*

Before stating the Proposition we need to introduce some additional auxiliary sequences:

Definition 12.

- *$\{M_j\}_{j \in \mathbb{N}_+}$ is the sequence of the sum of the first j work contributions, i.e.*

  $$M_0 = 0,\, M_j = \sum_{k=1}^{j} \Delta W_k\,. \tag{A.19}$$

  In this way the work of the optimal sparse grid (A.18) *corresponding to the first w indices is*

  $$W_{\mathcal{I}(w),m} = M_w \tag{A.20}$$

- $$\{\Delta \widetilde{E}_k\}_{k \in \mathbb{N}_+} = \underbrace{\Delta E_1,\, \Delta E_1,\, \Delta E_1 \ldots}_{\Delta W_1\, times}, \underbrace{\Delta E_2,\, \Delta E_2,\, \Delta E_2 \ldots}_{\Delta W_2\, times}\,, \tag{A.21}$$

  i.e. $\Delta \widetilde{E}_{M_{j-1}+s} = \Delta E_j, \quad s = 1, \ldots, \Delta W_j\,.$

- $$\{\widetilde{P}_k\}_{k \in \mathbb{N}_+} = \underbrace{\frac{\Delta E_1}{\Delta W_1}, \frac{\Delta E_1}{\Delta W_1}, \frac{\Delta E_1}{\Delta W_1} \ldots}_{\Delta W_1\, times}, \underbrace{\frac{\Delta E_2}{\Delta W_2}, \frac{\Delta E_2}{\Delta W_2}, \frac{\Delta E_2}{\Delta W_2} \ldots}_{\Delta W_2\, times}\,, \tag{A.22}$$

  i.e. $\widetilde{P}_{M_{j-1}+s} = P_j, \quad s = 1, \ldots, \Delta W_j\,.$

We will also need the following lemma:

Lemma 13. *Given a positive decreasing q-summable sequence $\{a_j\}_{j \in \mathbb{N}_+}$, for every $q > 0$ it holds*

$$\sup_{k \in \mathbb{N}_+} \left(k^{1/q} a_k\right) \leq \|a\|_{\ell^q}\,.$$

Proof. The thesis follows from the following chain of inequalities, that holds for all $k \in \mathbb{N}_+$ and for all $q > 0$,

$$k a_k^q \leq \sum_{j=1}^{k} a_j^q \leq \sum_{j=1}^{\infty} a_j^q = \|a\|_{\ell^q}^q\,.$$

□

We are now ready to state and proof the main result of this Section.

Theorem 14. *Consider the sparse grid* (A.11) *built using a family of nested interpolation points over the set* (A.18) *containg the w most profitable multiindices, and suppose that Assumption 11 holds. Consider the sequence $\{P_k\}_{k \in \mathbb{N}_+}$ of decreasing ordered profits and assume that it satisfies*

$$\left(\sum_{j>0} P_j^\tau \Delta W_j \right)^{1/\tau} < \infty \tag{A.23}$$

for some $0 < \tau < 1$. Then the sparse grid error decays at least as

$$\left\| u - \mathcal{S}^m_{\mathcal{I}(w)}[u] \right\|_{V \otimes L^2_\rho(\Gamma)} \leq \frac{\tau}{1-\tau} W^{1-1/\tau}_{\mathcal{I}(w),m} \left(\sum_{j>0} P_j^\tau \Delta W_j \right)^{1/\tau}. \tag{A.24}$$

Proof. For a generic sparse grid the following error decomposition holds:

$$\left\| u - \mathcal{S}^m_{\mathcal{I}(w)}[u] \right\|_{V \otimes L^2_\rho(\Gamma)} = \left\| \sum_{\mathbf{i} \notin \mathcal{I}(w)} \Delta^{m(\mathbf{i})}[u] \right\|_{V \otimes L^2_\rho(\Gamma)} \leq \sum_{\mathbf{i} \notin \mathcal{I}(w)} \left\| \Delta^{m(\mathbf{i})} u \right\|_{V \otimes L^2_\rho(\Gamma)} = \sum_{j>w} \Delta E_j. \tag{A.25}$$

Next we recast the previous sum of error contributions in terms of the auxiliary sequence $\widetilde{P}_k$,

$$\sum_{j>w} \Delta E_j = \sum_{j>w} \sum_{s=1}^{\Delta W_j} \frac{\Delta \widetilde{E}_{M_{j-1}+s}}{\Delta W_j} = \sum_{k>M_w} \widetilde{P}_k\,. \tag{A.26}$$

Observe that

$$\left(\sum_{k>0} \widetilde{P}_k^\tau \right)^{1/\tau} = \left(\sum_{j>0} P_j^\tau \Delta W_j \right)^{1/\tau}.$$

Thus, using Lemma 13 and the hypothesis (A.23) we get that $k^{1/\tau}\widetilde{P}_k$ is a bounded quantity. Indeed, there holds

$$\sup_{k \in \mathbb{N}_+} \left(k^{1/\tau} \widetilde{P}_k \right) \leq \left\| \widetilde{P}_k \right\|_{\ell^\tau} = \left(\sum_{k>0} \widetilde{P}_k^\tau \right)^{1/\tau} = \left(\sum_{j>0} P_j^\tau \Delta W_j \right)^{1/\tau} < \infty.$$

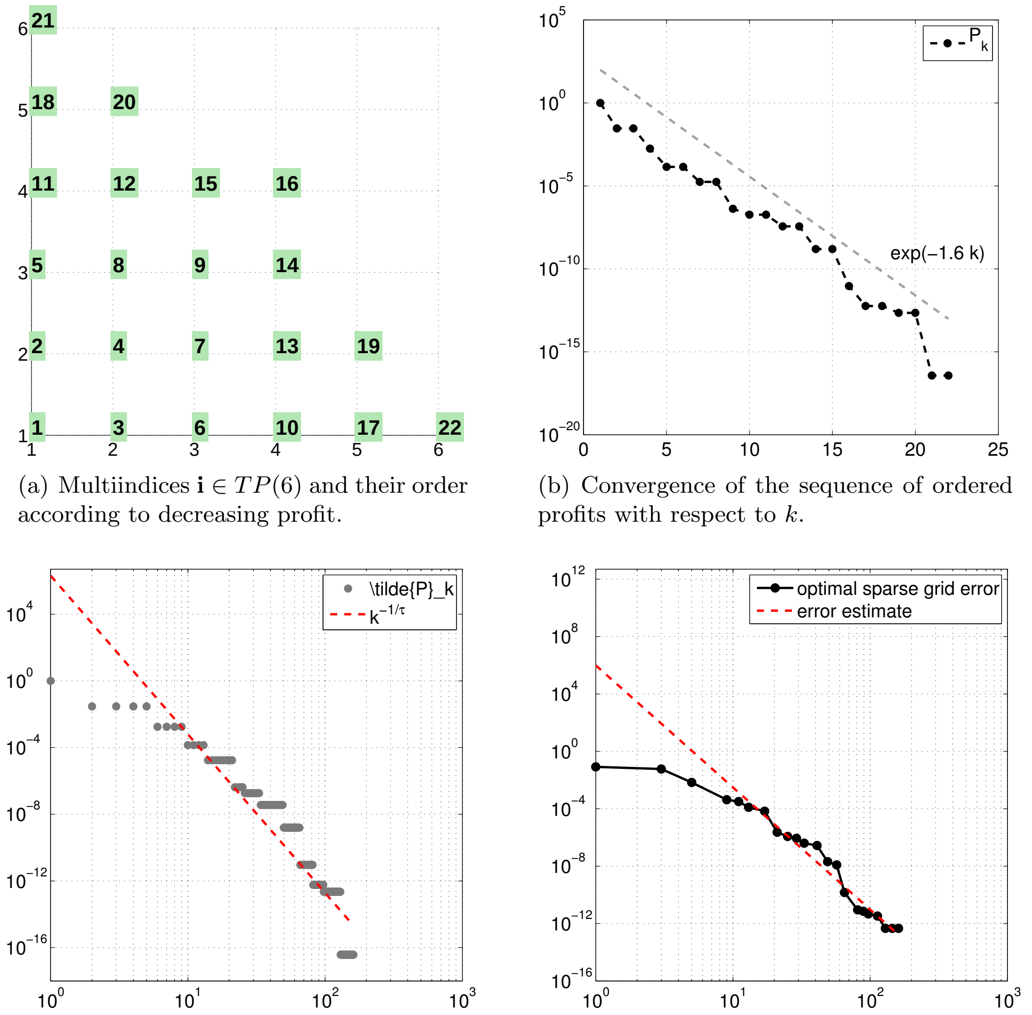
We can therefore write, combining (A.25) and (A.26),

$$\begin{aligned} \left\| u - \mathcal{S}^m_{\mathcal{I}(w)}[u] \right\|_{V \otimes L^2_\rho(\Gamma)} &\leq \sum_{k>M_w} \widetilde{P}_k = \sum_{k>M_w} k^{-1/\tau} k^{1/\tau} \widetilde{P}_k \\ &\leq \sup_{k \in \mathbb{N}_+} (k^{1/\tau} \widetilde{P}_k) \sum_{k>M_w} k^{-1/\tau} \leq \left(\sum_{j>0} P_j^\tau \Delta W_j \right)^{1/\tau} \sum_{k>M_w} k^{-1/\tau}. \end{aligned}$$

Finally, we bound

$$\sum_{k>M_w} k^{-1/\tau} \leq \int_{M_w}^\infty x^{-1/\tau} dx = \frac{\tau}{\tau - 1} M_w^{1-1/\tau},$$

and the proof is concluded by exploiting the fact that $M_w = W_{\mathcal{I}(w),m}$, see eq. (A.20). □



(a) Multiindices $\mathbf{i} \in TP(6)$ and their order according to decreasing profit.

(b) Convergence of the sequence of ordered profits with respect to k.

(c) Convergence of $\{\widetilde{P}_k\}_{k\in\mathbb{N}_+}$ and estimate of the rate τ.

(d) Comparison of the numerical error and of the predicted error decay according to equation (A.24).

Figure A.4: Plots for the optimal sparse grid convergence.

Remark 15. *We are indeed requiring that $\{P_k\}_{k\in\mathbb{N}_+}$ belongs to a weighted ℓ^τ space, whose weights are the work contributions of the hierarchical surpluses. The decay of the sparse grid error will be faster if τ is smaller: in particular, Theorem 14 guarantees convergence only if $\tau < 1$.*

Remark 16. *In the proof we have shown that* (A.23) *is indeed equivalent to the condition $\{\widetilde{P}_k\}_{k\in\mathbb{N}_+} \in \ell^\tau$. Note that $\{\widetilde{P}_k\}_{k\in\mathbb{N}_+}$ is "piecewise constant" by construction, and hence it will not have in general the same summability of the profit sequence $\{P_k\}_{k\in\mathbb{N}_+}$: equivalently, one may think of the terms of $\{\widetilde{P}_k\}_{k\in\mathbb{N}_+}$ as "decaying more slowly" than the terms of $\{P_k\}_{k\in\mathbb{N}_+}$.*

We now verify the accuracy of the convergence estimate (A.24). As in Chapter 3, we consider the model function $f = \frac{1}{1+c_1y_1+c_2y_2}$, with $c_1 = c_2 = 0.1$. For such function, we first compute the error and work contributions for the hierarchical surplus in $TP(6) = \{\mathbf{i} \in \mathbb{N}^2_+,\ \max\{i_1, i_2\} \leq 6\}$ and hence the profits, according to equations (A.14), (A.15), (A.16). The $L^2_\rho(\Gamma)$ norm needed to compute the error contribution is approximated with a very accurate sparse grid, and we interrupt the computation when the profits reach 10^{-17}.

Next, we order the computed profits in decreasing order: note that the induced sequence $\mathcal{I}(w)$ satisfies the admissibility condition, see figure A.4(a). The resulting sequence $\{P_k\}_{k\in\mathbb{N}_+}$ seems to converge exponentially, see in Figure A.4(b). Next, we build $\{\widetilde{P}_k\}_{k\in\mathbb{N}_+}$. We can then compute numerically a crude estimate for τ e.g. by estimating with a least square approach a value for τ such that $\sup\left(k^{1/\tau}\widetilde{P}_k\right)$ is bounded, i.e. $\widetilde{P}_k \sim k^{1/\tau}$, see Figure A.4(c). The estimated value for τ is $\tau \approx 0.10$. Such τ safisties condition (A.23). The quantity $\sum_{j>0} P_j^\tau \Delta W_j$ is indeed bounded (its numerical value is ≈ 21), and thus τ can be used to estimate the convergence rate according to equation (A.24), which results in approximately 8.50.

Finally, we actually build the sequence of sparse grid using the sequence of sets $\mathcal{I}(w)$ and compare the resulting $L^2_\rho(\Gamma)$ error with the predicted convergence rate. Results are shown in Figure A.4(d) and suggest that the theoretical rate is quite sharp.

1.2.a The inclusion case

In this section we consider again the inclusions problem, and we apply Theorem 14 to determine the convergence rate of the quasi-optimal sparse grid for this problem, built taking multi-indices in descending order according to the estimated profits. To this end, we need to detail the estimates for the error and the work contribution for the problem at hand, verify Assumption 11 on the admissibility of the optimal sets and assess the weighted ℓ^τ summability (A.23) of the profits. We will consider a sparse grid built on Clenshaw–Curtis points, whose Lebesgue constant $\mathbb{L}(i)$ satisfies

$$\mathbb{L}(i) \leq \frac{2}{\pi}\log(i+1)+1, \tag{A.27}$$

using the relation $m(i)$ as in equation (A.12) (see e.g. [28, 29] for the Lebesgue constant of the Clenshaw–Curtis points). We recall that the following estimate holds for the interpolant operator $\mathcal{U}^i$ defined in (A.11):

$$\left\|\mathcal{U}^i[f]\right\|_{L^\infty(\Gamma_n)} \leq \mathbb{L}(i)\left\|f\right\|_{L^\infty(\Gamma_n)}, \quad f\in\mathcal{C}^0(\Gamma_n). \tag{A.28}$$

As a first step, we need to derive explicit estimates for the work and error contribution. Analogously to what was done in Section 3.4.a, the work contribution of each hierarchical surplus can be computed as

$$\Delta W(\mathbf{i}) = W_{\{\mathcal{J}\cup\mathbf{i}\},m} - W_{\mathcal{J},m} = \prod_{n=1}^N m(i_n) - m(i_n-1) = \prod_{n=1}^N r(i_n). \tag{A.29}$$

As for the error estimate, we use the same estimate as in Section 3.4.a, see equation (3.33), which can be proved rigorously for the case of the inclusions problem. We thus have the following Lemma:

Lemma 17. *For the inclusions problem there holds*

$$\Delta E(\mathbf{i}) \leq Ce^{-\sum_{n=1}^N \widehat{g}_n m(i_n-1)}\prod_{n=1}^N \mathbb{L}_n^{m(i_n)}. \tag{A.30}$$

Proof. We start by considering again the argument in Section 4.5.c:

$$\begin{aligned}\Delta E(\mathbf{i}) &= \left\|\Delta^{m(\mathbf{i})}[u]\right\|_{H^1_0(D)\otimes L^2_\rho(\Gamma)} = \left\|\Delta^{m(\mathbf{i})}[\sum_{\mathbf{p}\in\mathbb{N}^N} u_\mathbf{p}\mathcal{L}_\mathbf{p}]\right\|_{H^1_0(D)\otimes L^2_\rho(\Gamma)}\\ &= \left\|\sum_{\mathbf{p}\in\mathbb{N}^N} u_\mathbf{p}\Delta^{m(\mathbf{i})}[\mathcal{L}_\mathbf{p}]\right\|_{H^1_0(D)\otimes L^2_\rho(\Gamma)} = \left\|\sum_{\mathbf{p}\geq m(\mathbf{i}-\mathbf{1})} u_\mathbf{p}\Delta^{m(\mathbf{i})}[\mathcal{L}_\mathbf{p}]\right\|_{H^1_0(D)\otimes L^2_\rho(\Gamma)}\\ &\leq \sum_{\mathbf{p}\geq m(\mathbf{i}-\mathbf{1})} \left\|u_\mathbf{p}\right\|_{H^1_0(D)}\left\|\Delta^{m(\mathbf{i})}[\mathcal{L}_\mathbf{p}]\right\|_{L^2_\rho(\Gamma)}.\end{aligned}$$

Next, using (A.28) and the defintion (A.11) of $\Delta^{m(\mathbf{i})}$ we bound the norm of the hierarchical surplus applied to the Legendre polynomials as

$$\left\|\Delta^{m(\mathbf{i})}[\mathcal{L}_{\mathbf{p}}]\right\|_{L^2_\rho(\Gamma)} = \prod_{n=1}^{N} \left\|\Delta^{m(i_n)}[L_{p_n}]\right\|_{L^2_{\rho_n}(\Gamma_n)} \leq \prod_{n=1}^{N} \left\|\Delta^{m(i_n)}[\sqrt{2p_n+1}\mathscr{L}_{p_n}]\right\|_{L^\infty(\Gamma_n)}$$
$$\leq \prod_{n=1}^{N} 2\,\mathbb{L}_n^{m(i_n)}\sqrt{2p_n+1}\ \|\mathscr{L}_{p_n}\|_{L^\infty(\Gamma_n)} = \prod_{n=1}^{N} 2\,\mathbb{L}_n^{m(i_n)}\sqrt{2p_n+1}.$$

Recalling estimate (A.4) for the decay of the Legendre coeffcients of the inclusions problem, one obtains

$$\Delta E(\mathbf{i}) \leq \sum_{\mathbf{p}\geq m(\mathbf{i}-\mathbf{1})} \|u_{\mathbf{p}}\|_{H^1(D)} \prod_{n=1}^{N} 2\sqrt{2p_n+1}\,\mathbb{L}_n^{m(i_n)} \leq \sum_{\mathbf{p}\geq m(\mathbf{i}-\mathbf{1})} C_{Leg} \prod_{n=1}^{N} 2e^{-g_n p_n}(2p_n+1)\mathbb{L}_n^{m(i_n)}$$
$$\leq C_{Leg} \prod_{n=1}^{N} \sum_{p_n\geq m(i_n-1)} 2(2p_n+1)e^{-g_n p_n}\mathbb{L}_n^{m(i_n)}$$
$$\leq C_{Leg} \prod_{n=1}^{N} \mathbb{L}_n^{m(i_n)} \left(4 \sum_{p_n\geq m(i_n-1)} e^{-g_n p_n} p_n + 2 \sum_{p_n\geq m(i_n-1)} e^{-g_n p_n}\right)$$
$$\leq \widehat{C}_{Leg} \prod_{n=1}^{N} \mathbb{L}_n^{m(i_n)} e^{-\widehat{g}_n m(i_n-1)},$$

for some $\widehat{g}_n < g_n$ and $\widehat{C}_{Leg} > C_{Leg}$ (as in Remark 4). □

Thus, to estimate the ℓ^τ weighted summability (A.23) of the profits, we are led to investigate the summability of the sum

$$\sum_{\mathbf{i}\in\mathbb{N}^N} \prod_{n=1}^{N} \left[\left(\frac{e^{-\widehat{g}_n m(i_n-1)}\mathbb{L}_n(m(i_n))}{r(i_n)}\right)^\tau r(i_n)\right], \tag{A.31}$$

for which we can prove the following proposition.

Lemma 18. *The series* (A.31) *is finite for every* $\tau < 1$.

Proof. We start by observing that since the general term of (A.31) is a product we can actually write as

$$\sum_{\mathbf{i}\in\mathbb{N}^N} \prod_{n=1}^{N} \left[\left(\frac{e^{-\widehat{g}_n m(i_n-1)}\mathbb{L}_n(m(i_n))}{r(i_n)}\right)^\tau r(i_n)\right] = \prod_{n=1}^{N} \sum_{i_n=0}^{\infty} \left(\frac{e^{-\widehat{g}_n m(i_n-1)}\mathbb{L}_n(m(i_n))}{r(i_n)}\right)^\tau r(i_n),$$

so that indeed we only need to study the summability of

$$\sum_{i_n=0}^{\infty} \left(\frac{e^{-\widehat{g}_n m(i_n-1)}\mathbb{L}_n(m(i_n))}{r(i_n)}\right)^\tau r(i_n). \tag{A.32}$$

First, we bound the Lebesgue constant. It is enough for our purposes to use the crude bound

$$\mathbb{L}_n(m(i)) = \frac{2}{\pi}\log(m(i)+1)+1 \leq \frac{2}{\pi}m(i)+1 \leq 2m(i)+1 \leq 2(m(i)+1).$$

Next, observe that for $i > 2$ it holds

$$\frac{2(m(i)+1)}{r(i)} = \frac{2(2^{i-1}+1+1)}{2^{i-2}} = \frac{2^{i-1}+2}{2^{i-3}} \leq \frac{2^i}{2^{i-3}} \leq 8.$$

We can then bound the generic term of (A.32) for $i > 2$ as

$$\left(\frac{e^{-gm(i-1)}\mathbb{L}_n(m(i))}{r(i)}\right)^\tau r(i) \leq 8^\tau e^{-\tau g m(i-1)} r(i) = 8^\tau e^{-\tau g(2^{i-2}+1)} 2^{i-2}.$$

We can now prove that (A.32) is finite. Indeed, using summation by parts one obtains

$$\begin{aligned}\sum_{i=2}^{\infty} 8^\tau e^{-\tau g(2^{i-2}+1)} 2^{i-2} &\leq 8^\tau e^{-\tau g} \sum_{p=1}^{\infty} e^{-\tau g p} p \\ &= \underbrace{\frac{8^\tau e^{-\tau g}}{2} \frac{1}{e^{\tau g}-1}}_{C} \sum_{p=1}^{\infty} \left(e^{-\tau g(p-1)} - e^{-\tau g p}\right) p \\ &= C\left(e^{-\tau g} + \sum_{p=1}^{\infty} e^{-\tau g p}\right) = C\left(e^{-\tau g} + \frac{e^{-\tau g}}{1-e^{-\tau g}}\right) < +\infty.\end{aligned}$$

□

Finally, we verify Assumption 11, i.e. that the sequence of multi-indices ordered by decreasing profit is admissible.

Lemma 19. *Let $\mathbf{e}_j$ be the j-th canonical vector, $j = 1, \ldots, N$. Then, the sequence $\mathcal{I}(w)$ of multi-indices ordered by decreasing profits defined in* (A.18)*, with $\Delta W(\mathbf{i})$ as in* (A.29) *and $\Delta E(\mathbf{i})$ as in* (A.30) *is such that*

$$P(\mathbf{i}+\mathbf{e}_j) \leq P(\mathbf{i}), \quad \forall\, \mathbf{i} \in \mathbb{N}_+^N, \quad \forall j = 1, \ldots, N,$$

i.e. $\mathcal{I}(w)$ is admissible, provided that for $\widehat{g}_j$ in (A.30) *it holds*

$$\widehat{g}_j \geq \frac{1}{2}\log\left(\frac{\frac{2}{\pi}\log 6 + 1}{\frac{2}{\pi}\log 4 + 1}\right) \approx 0.06, \qquad \textit{for } j = 1, \ldots, N. \tag{A.33}$$

Proof. From the definition of profits (A.16), the admissibility condition $P(\mathbf{i}+\mathbf{e}_j) \leq P(\mathbf{i})$ is equivalent to

$$\frac{\Delta E(\mathbf{i}+\mathbf{e}_j)}{\Delta E(\mathbf{i})} \leq \frac{\Delta W(\mathbf{i}+\mathbf{e}_j)}{\Delta W(\mathbf{i})}.$$

Given the expression for ΔW, ΔE in (A.29) and (A.30), this is actually equivalent to

$$\frac{e^{-\widehat{g}m(i_j)}\mathbb{L}(m(i_j+1))}{e^{-\widehat{g}m(i_j-1)}\mathbb{L}(m(i_j))} \leq \frac{m(i_j+1)-m(i_j)}{m(i_j)-m(i_j-1)}. \tag{A.34}$$

We now insert the expression of (A.12) in the right hand side. This results in

$$\overline{m}(j) = \frac{m(i_j+1)-m(i_j)}{m(i_j)-m(i_j-1)} = \begin{cases} (i_j \geq 3) & = \dfrac{2^{i_j}-2^{i_j-1}}{2^{i_j-1}-2^{i_j-2}} = \dfrac{2^{i_j-1}}{2^{i_j-2}} = 2 \\ (i_j = 2) & = \dfrac{m(3)-m(2)}{m(2)-m(1)} = \dfrac{5-3}{3-1} = 1 \\ (i_j = 1) & = \dfrac{m(2)-m(1)}{m(1)-m(0)} = \dfrac{3-1}{1-0} = 2. \end{cases}$$

so that (A.34) is equivalent to

$$e^{-\widehat{g}(m(i_j)-m(i_j-1)}\frac{\mathbb{L}(m(i_j+1))}{\mathbb{L}(m(i_j))} \leq \overline{m}(j). \tag{A.35}$$

We now verify (A.35) for the three different cases $i_j = 1, i_j = 2, i_j \geq 3$.

Case $i_j = 1$. Recalling equation (A.27) for the Lebesgue constant and equation (A.12) for $m(i)$ there holds

$$\frac{\mathbb{L}(m(i_j+1))}{\mathbb{L}(m(i_j))} = \frac{\mathbb{L}(m(2))}{\mathbb{L}(m(1))} = \frac{\mathbb{L}(3)}{\mathbb{L}(1)} = \frac{\frac{2}{\pi}\log(4)+1}{\frac{2}{\pi}\log(2)+1} \approx 1.31.$$

Next, observe that $e^{-\widehat{g}(m(i_j)-m(i_j-1)} < 1$. Therefore (A.35) holds.

Case $i_j = 2$. Similarly to the previous case, it holds

$$\frac{\mathbb{L}(m(i_j+1))}{\mathbb{L}(m(i_j))} = \frac{\mathbb{L}(m(3))}{\mathbb{L}(m(2))} = \frac{\mathbb{L}(5)}{\mathbb{L}(3)} = \frac{\frac{2}{\pi}\log(6)+1}{\frac{2}{\pi}\log(4)+1} \approx 1.13.$$

Equation (A.35) is therefore equivalent to

$$e^{-2g_j}\frac{\frac{2}{\pi}\log(6)+1}{\frac{2}{\pi}\log(4)+1} \leq 1,$$

hence condition (A.33).

Case $i_j \geq 3$. There holds

$$\begin{aligned}\frac{\mathbb{L}(m(i_j+1))}{\mathbb{L}(m(i_j))} &= \frac{\frac{2}{\pi}\log(m(i_j+1)+1)+1}{\frac{2}{\pi}\log(m(i_j)+1)+1} = \frac{\log(m(i_j+1)+1)+\frac{\pi}{2}}{\log(m(i_j)+1)+\frac{\pi}{2}} = \frac{\log(2^{i_j}+2)+\frac{\pi}{2}}{\log(2^{i_j-1}+2)+\frac{\pi}{2}} \\ &\leq \frac{\log(2^{i_j+1})+\frac{\pi}{2}}{\log(2^{i_j-1})+\frac{\pi}{2}} = \frac{(i_j+1)\log(2)+\frac{\pi}{2}}{(i_j-1)\log(2)+\frac{\pi}{2}}.\end{aligned}$$

Being this function decreasing in i_j, it is easy to show that $\dfrac{\mathbb{L}(m(i_j+1))}{\mathbb{L}(m(i_j))} \leq 2$, $\forall\, i_j \geq 3$.

□

Remark 20. $\widehat{g}$ *in* (A.33) *is related to the size of the analytic continuation region of the solution* u *of the inclusion problem (see Lemma 2 and Prop. 3). This is in turn related to the support of the random variables* y_i *defining the diffusion coefficients inside the inclusions, see Prop. 9 and Theorem 10.*

Remark 21. *A closer look at the proof of Lemma 19 reveals that the only case where condition* (A.33) *plays a role is for* $i_j = 2$*. This means that the non-admissibility of the set cannot take place outside the hypercube* $\{\mathbf{i} \in \mathbb{N}_+^N : i_j \geq 3, j = \ldots, N\}$*. Therefore, in the asymptotic limit the sets of the most profitable indices are always admissible,* $\forall\, \widehat{g} > 0$*.*

Collecting results in Lemma 17, Lemma 18 and Lemma 19, we are now in position to apply Theorem 14 and state the final result on the convergence of the optimal sparse grid method for the inclusion problem.

Theorem 22. *For a sufficiently large level w, the quasi-optimal sparse grid for the inclusion problem converges with rate $1 - 1/\tau$ for any $\tau < 1$. Under condition* (A.33)*, the result holds $\forall\, w > 0$.*

Remark 23. *The Theorem just stated implies that the convergence of the optimal sparse grid in the case of the inclusions problem is more than algebraic. This is consistent with the numerical results obtained in Section 2.4, where a more than algebraic convergence is observed already for the non optimized standard SM grid with Clenshaw–Curtis abscissae (see e.g. figure 2.3, orange line).*

Bibliography

[1] G. E. B. Archer, A. Saltelli, and I. M. Sobol, Sensitivity measures, anova-like techniques and the use of bootstrap, *Journal of Statistical Computation and Simulation* **58** (1997) 99–120.

[2] K. I. Babenko, Approximation by trigonometric polynomials in a certain class of periodic functions of several variables, *Soviet Math. Dokl.* **1** (1960) 672–675.

[3] I. Babuška, R. Tempone, and G. E. Zouraris, Galerkin finite element approximations of stochastic elliptic partial differential equations, *SIAM J. Numer. Anal.* **42** (2004) 800–825.

[4] 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** (2007) 1005–1034.

[5] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone, Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison, *Spectral and High Order Methods for Partial Differential Equations* J.S. Hesthaven and E.M. Ronquist eds., Lecture Notes in Computational Science and Engineering **76** (Springer, 2011) 43–62, Selected papers from the ICOSAHOM '09 conference, June 22-26, Trondheim, Norway.

[6] T. Bagby, L. Bos, and N. Levenberg, Multivariate simultaneous approximation, *Constr. Approx.* **18** (2002) 569–577.

[7] V. Barthelmann, E. Novak, and K. Ritter, High dimensional polynomial interpolation on sparse grids, *Adv. Comput. Math.* **12** (2000) 273–288.

[8] J. Bear, *Dynamics of fluids in porous media* (American Elsevier Pub. Co., 1972).

[9] J. Beck, F. Nobile, L. Tamellini, and R. Tempone, On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods, To appear on Mathematical Models & Methods in Applied Sciences. Also available as MOX-Report 23-2011.

[10] J. Beck, F. Nobile, L. Tamellini, and R. Tempone, Implementation of optimal Galerkin and Collocation approximations of PDEs with Random Coefficients, *ESAIM: Proc.* **33** (2011) 10–21.

[11] M. Bieri, R. Andreev, and C. Schwab, Sparse tensor discretization of elliptic spdes, *SAM-Report*, 2009-07, Seminar für Angewandte Mathematik, ETH, Zurich, 2009.

[12] K. Bjorlykke and K. Hoeg, Effects of burial diagenesis on stresses, compaction and fluid flow in sedimentary basins, *Mar. and Petrol. Geol.* **14** (1997) 267–276.

[13] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer Series in Computational Mathematics **15** (Springer-Verlag, 1991).

[14] A. T. Buller, P. A. Bjorkum, P. Nadeau, and O. Walderhaug, Distribution of hydrocarbons in sedimentary basins, *Statoil ASA, Res and Techn. Memoir* **7** (2005) 1–15.

[15] H.J Bungartz and M. Griebel, Sparse grids, *Acta Numer.* **13** (2004) 147–269.

[16] R. E. Caflisch, Monte Carlo and quasi-Monte Carlo methods, Acta numerica, 1998, Acta Numer. **7** (Cambridge Univ. Press, Cambridge, 1998) 1–49.

[17] C. Canuto, M.Y. Hussaini, A. Quateroni, and T.A. Zang, *Spectral methods in fluid dynamics* (Springer-Verlag, 1988).

[18] C. Canuto and T. Kozubek, A fictitious domain approach to the numerical solution of PDEs in stochastic domains, *Numer. Math.* **107** (2007) 257–293.

[19] J. Charrier, Strong and weak error estimates for the solutions of elliptic partial differential equations with random coefficients, *INRIA - Rapport de recherche*, 7300-version 3, Institut National de recherche en informatique et en automatique - INRIA, 2011.

[20] A. Cohen, R. DeVore, and C. Schwab, Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs, *SAM-Report*, 2010-03, Seminar für Angewandte Mathematik, ETH, Zurich, 2010.

[21] A. Cohen, R. DeVore, and C. Schwab, Convergence rates of best n-term Galerkin approximations for a class of elliptic sPDEs, *Foundations of Computational Mathematics* **10** (2010) 615–646.

[22] P.G. Constantine, M.S. Eldred, and E.T. Phipps, Sparse pseudospectral approximation method, *ArXiv repository*, item number: arXiv:1109.2936v1, 2011.

[23] G. Dagan, U. Hornung, and P. Knabner, *Mathematical modeling for flow and transport through porous media* (Kluwer Academic Publishers, 1991).

[24] G. Dagan and S.P. Neuman, *Subsurface flow and transport: A stochastic approach*, International Hydrology Series (Cambridge University Press, 2005).

[25] P.J. Davis, *Interpolation and approximation* (Dover Publications Inc., 1975).

[26] J. Dick and F. Pillichshammer, *Digital nets and sequences: Discrepancy theory and quasi-monte carlo integration* (Cambridge University Press, 2010).

[27] A. Doostan, R. G. Ghanem, and J. Red-Horse, Stochastic model reduction for chaos representations, *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 3951–3966.

[28] V. K. Dzjadik and V. V. Ivanov, On asymptotics and estimates for the uniform norms of the Lagrange interpolation polynomials corresponding to the Chebyshev nodal points, *Anal. Math.* **9** (1983) 85–97.

[29] H. Ehlich and K. Zeller, Auswertung der Normen von Interpolationsoperatoren, *Math. Ann.* **164** (1966) 105–112.

[30] H. C. Elman, C. W. Miller, E. T. Phipps, and R. S. Tuminaro, Assessment of Collocation and Galerkin approaches to linear diffusion equations with random data, *International Journal for Uncertainty Quantification* **1** (2011) 19–33.

[31] O. G. Ernst, A. Mugler, H.-J. Starkloff, and E. Ullmann, On the convergence of generalized polynomial chaos expansions, *ESAIM: Mathematical Modelling and Numerical Analysis* **46** (2012) 317–339.

[32] O. G. Ernst, C. E. Powell, D. J. Silvester, and E. Ullmann, Efficient solvers for a linear stochastic Galerkin mixed formulation of diffusion problems with random data, *SIAM J. Sci. Comput.* **31** (2008/09) 1424–1447.

[33] X. Fernique, Intégrabilité des vecteurs gaussiens, *C. R. Acad. Sci. Paris Sér. A-B* **270** (1970) A1698–A1699.

[34] L. Formaggia, A Guadagnini, I. Imperiali, G. Porta, M. Riva, A. Scotti, and L. Tamellini, A numerical model for the geological compaction of sedimentary basins with sensitivity analysis, In preparation.

[35] S. Franzetti and A. Guadagnini, Probabilistic estimation of well catchments in heterogeneous aquifers, *Journal of Hydrology* **174** (1996) 149–171.

[36] P. Frauenfelder, C. Schwab, and R. A. Todor, Finite elements for elliptic problems with stochastic coefficients, *Comput. Methods Appl. Mech. Engrg.* **194** (2005) 205–228.

[37] J. Galvis and M. Sarkis, Approximating infinity-dimensional stochastic Darcy's equations without uniform ellipticity, *SIAM J. Numer. Anal.* **47** (2009) 3624–3651.

[38] B. Ganapathysubramanian and N. Zabaras, Sparse grid collocation schemes for stochastic natural convection problems, *Journal of Computational Physics* **225** (2007) 652–685.

[39] W. Gautschi, *Orthogonal polynomials: Computation and approximation* (Oxford University Press, 2004).

[40] A. Genz and B. D. Keister, Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight, *J. Comput. Appl. Math.* **71** (1996) 299–309.

[41] T. Gerstner and M. Griebel, Dimension-adaptive tensor-product quadrature, *Computing* **71** (2003) 65–87.

[42] R. G. Ghanem and P. D. Spanos, *Stochastic finite elements: a spectral approach* (Springer-Verlag, 1991).

[43] C. J. Gittelson, Stochastic Galerkin discretization of the log-normal isotropic diffusion problem, *Math. Models Methods Appl. Sci.* **20** (2010) 237–263.

[44] C.J. Gittelson, An adaptive stochastic galerkin method, *SAM-Report*, 2011-11, Seminar für Angewandte Mathematik, ETH, Zurich, 2011.

[45] M. Griebel and S. Knapek, Optimized general sparse grid approximation spaces for operator equations, *Math. Comp.* **78** (2009) 2223–2257.

[46] H. Harbrecht, On output functionals of boundary value problems on stochastic domains, *Math. Methods Appl. Sci.* **33** (2010) 91–102.

[47] H. Harbrecht, R. Schneider, and C. Schwab, Sparse second moment analysis for elliptic problems in stochastic domains, *Numer. Math.* **109** (2008) 385–414.

[48] F. Heiss and V. Winschel, Likelihood approximation by numerical integration on sparse grids, *J. Econometrics* **144** (2008) 62–80.

[49] V. H. Hoang and C. Schwab, N-term galerkin wiener chaos approximations of elliptic pdes with lognormal gaussian random inputs, *SAM-Report*, 2011-59, Seminar für Angewandte Mathematik, ETH, Zurich, 2011.

[50] F. P. Incropera, T. L. Bergman, A. S. Lavine, and D. P. DeWitt, *Fundamentals of heat and mass transfer* (John Wiley & Sons, 2011).

[51] B. Jansson, *Random number generators* (Almqvist & Wiksell, 1966).

[52] S. Joe and F. Y. Kuo, Constructing Sobol' sequences with better two-dimensional projections, *SIAM J. Sci. Comput.* **30** (2008) 2635–2654.

[53] R.A. Johnson and D.W. Wichern, *Applied multivariate statistical analysis* (Pearson Prentice Hall, 2007).

[54] I. T. Jolliffe, *Principal component analysis*, second ed., Springer Series in Statistics (Springer-Verlag, 2002).

[55] D. R. Jones, A taxonomy of global optimization methods based on response surfaces, *J. Global Optim.* **21** (2001) 345–383.

[56] A. Klimke, *Uncertainty modeling using fuzzy arithmetic and sparse grids*, Ph.D. thesis, Universität Stuttgart, Shaker Verlag, Aachen, 2006.

[57] D. E. Knuth, *The art of computer programming. Vol. 2: Seminumerical algorithms* (Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont, 1969).

[58] A. S. Kronrod, *Nodes and weights of quadrature formulas. Sixteen-place tables*, Authorized translation from the Russian (Consultants Bureau, 1965).

[59] R. H. Lander and O. Walderhaug, Predicting porosity through simulating sandstone compaction and quartz cementation, *AAPG Bulletin* **83(3)** (1999).

[60] O. P. Le Maître and O. M. Knio, *Spectral methods for uncertainty quantification*, Scientific Computation (Springer, 2010).

[61] O. P. Le Maître, H. N. Najm, R. G. Ghanem, and O. M. Knio, Multi-resolution analysis of Wiener-type uncertainty propagation schemes, *J. Comput. Phys.* **197** (2004) 502–531.

[62] P. Lévy, *Processus stochastiques et mouvement brownien*, Les Grands Classiques Gauthier-Villars. [Gauthier-Villars Great Classics] (Éditions Jacques Gabay, 1992).

[63] M. Loève, *Probability theory. I*, fourth ed. (Springer-Verlag, 1977).

[64] M. Loève, *Probability theory. II*, fourth ed. (Springer-Verlag, 1978).

[65] Wei-Liem Loh, On Latin hypercube sampling, *Ann. Statist.* **24** (1996) 2058–2080.

[66] H. G. Matthies and A. Keese, Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations, *Comput. Methods Appl. Mech. Engrg.* **194** (2005) 1295–1331.

[67] K. L. Milliken, Late diagenesis and mass transfer in sandstone-shale sequences, *Treatis on Geochemistry* **7** (2004) 115–158.

[68] D.C. Montgomery, *Design and analysis of experiments* (Wiley, 2008).

[69] R. B. Nelsen, *An introduction to copulas*, second ed., Springer Series in Statistics (Springer, 2006).

[70] H. Niederreiter, *Random number generation and quasi-monte carlo methods*, CBMS-NSF regional conference series in applied mathematics (Society for Industrial and Applied Mathematics, 1992).

[71] F. Nobile and R. Tempone, Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients, *Internat. J. Numer. Methods Engrg.* **80** (2009) 979–1006.

[72] 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** (2008) 2411–2442.

[73] 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** (2008) 2309–2345.

[74] A. Nouy, Generalized spectral decomposition method for solving stochastic finite element equations: invariant subspace problem and dedicated algorithms, *Comput. Methods Appl. Mech. Engrg.* (2007).

[75] A. Nouy, A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations, *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 4521–4537.

[76] A. Nouy, Méthode de construction de bases spectrales généralisées pour l'approximation de problèmes stochastiques, *Mécanique & Industries* **8** (2007) 283–288.

[77] A. Nouy, A. Clément, F. Schoefs, and N. Moës, An extended stochastic finite element method for solving stochastic partial differential equations on random domains, *Comput. Methods Appl. Mech. Engrg.* **197** (2008) 4663–4682.

[78] A. Nouy and O. P. Le Maître, Generalized spectral decomposition for stochastic nonlinear problems, *J. Comput. Phys.* **228** (2009) 202–235.

[79] E. H. Oelkers, P. A. Bjorkum, and W. M. Murphy, The mecanism of porosity reduction, stylolite development and quartz cementation in North Sea sandstone, Water-Rock Interaction (, 1992) 1183–1186.

[80] B. Øksendal, *Stochastic differential equations*, sixth ed., Universitext (Springer-Verlag, 2003).

[81] T. N. L. Patterson, The optimum addition of points to quadrature formulae, *Math. Comp. 22 (1968), 847–856; addendum, ibid.* **22** (1968) C1–C11.

[82] M. F. Pellissetti and R. G. Ghanem, Iterative solution of systems of linear equations arising in the context of stochastic finite elements, *Adv. Eng. Software* **31** (2000) 607–616.

[83] C. E. Powell and H. C. Elman, Block-diagonal preconditioning for spectral stochastic finite-element systems, *IMA J. Numer. Anal.* **29** (2009) 350–375.

[84] Catherine E. Powell and Elisabeth Ullmann, Preconditioning stochastic Galerkin saddle point systems, *SIAM J. Matrix Anal. Appl.* **31** (2010) 2813–2840.

[85] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, second ed., Texts in Applied Mathematics **37** (Springer-Verlag, 2007).

[86] P. A. Raviart and J. M. Thomas, A mixed finite element method for 2nd order elliptic problems, Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), Lecture Notes in Math. **606** (Springer, Berlin, 1977) 292–315.

[87] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, second ed., Springer Texts in Statistics (Springer-Verlag, 2004).

[88] S. M. Ross, *Simulation* (Academic Press, 2002).

[89] S. M. Ross, *A first course in probability* (Pearson Prentice Hall, 2010).

[90] Eveline Rosseel and Stefan Vandewalle, Iterative solvers for the stochastic finite element method, *SIAM J. Sci. Comput.* **32** (2010) 372–397.

[91] B. B. Sageman and T. W. Lyons, Geochemistry of fine-grained sediments and sedimentary rocks, *Treatise on Geochemistry* **7** (2004) 115–158.

[92] A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, Sensitivity analysis practices: strategies for model based inference, *Reliab. Eng. Syst. Saf.* **91** (2006) 1109–1125.

[93] J. Shen and L-L. Wang, Sparse spectral approximations of high-dimensional problems based on hyperbolic cross, submitted to SIAM J. Numer. Anal.

[94] I. H. Sloan and H. Woźniakowski, When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals?, *J. Complexity* **14** (1998) 1–33.

[95] S.A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, *Dokl. Akad. Nauk SSSR* **4** (1963) 240–243.

[96] I. M. Sobol', Sensitivity estimates for nonlinear mathematical models, *Math. Modeling Comput. Experiment* **1** (1993) 407–414 (1995).

[97] G. Stefanou, The stochastic finite element method: past, present and future, *Comput. Methods Appl. Mech. Engrg.* **198** (2009) 1031–1051.

[98] Michael Stein, Large sample properties of simulations using Latin hypercube sampling, *Technometrics* **29** (1987) 143–151.

[99] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, *Reliability Engineering & System Safety* **93** (2008) 964 – 979.

[100] G. Szegö, *Orthogonal polynomials*, Colloquium Publications - American Mathematical Society (American Mathematical Society, 1939).

[101] R. A. Todor and C. Schwab, Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients, *IMA J Numer Anal* **27** (2007) 232–261.

[102] L. N. Trefethen, Is Gauss quadrature better than Clenshaw-Curtis?, *SIAM Rev.* **50** (2008) 67–87.

[103] E. Ullmann, A Kronecker product preconditioner for stochastic Galerkin finite element discretizations, *SIAM J. Sci. Comput.* **32** (2010) 923–946.

[104] E. Ullmann, H. C. Elman, and Ernst O. G., Efficient iterative solvers for stochastic galerkin discretizations of log-transformed random diffusion problems, *DFG-SPP 1324 Preprint*, 95, DFG-Deutsche Forschungsgemeinschaft, 2011.

[105] O. Walderhaug, Precipitation rates for quartz cement in sandstones determined by fluid-inclusion microthermometry and temperature-history modelling, *J. Sed. Resear.* **A64** (1994) 324–333.

[106] O. Walderhaug, Kinetic modeling of quartz cementation and porosity loss in deeply buried sandstone reservoirs, *AAPG Bulletin* **80** (1996) 731–745.

[107] X. Wan and G. E. Karniadakis, An adaptive multi-element generalized polynomial chaos method for stochastic differential equations, *J. Comput. Phys.* **209** (2005) 617–642.

[108] X. Wan and G. E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, *SIAM J. Sci. Comput.* **28** (2006) 901–928.

[109] M. Wangen, *Physical principles of sedimentary basin analysis* (Cambridge University Press, 2010).

[110] D. Xiu and J.S. Hesthaven, High-order collocation methods for differential equations with random inputs, *SIAM J. Sci. Comput.* **27** (2005) 1118–1139.

[111] D. Xiu and G.E. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations, *SIAM J. Sci. Comput.* **24** (2002) 619–644.

[112] D. Xiu and D. M. Tartakovsky, Numerical methods for differential equations in random domains, *SIAM J. Sci. Comput.* **28** (2006) 1167–1185 (electronic).