Sparse tensor discretizations of elliptic PDEs with random input data

A dissertation submitted to
ETH Zürich

for the degree of
Doctor of Sciences

presented by
MARCEL BIERI
Dipl. Math. ETH
born January 7, 1979
citizen of Schangnau BE, Switzerland

accepted on the recommendation of
Prof. Dr. Christoph Schwab, ETH Zürich, examiner
Prof. Dr. Hermann G. Matthies, TU Braunschweig, co-examiner
Prof. Dr. Jan S. Hesthaven, Brown University, co-examiner

2009
Acknowledgments

First and foremost I would like to express my gratitude to my advisor Prof. Christoph Schwab, for his guidance during the last four years. He was always accessible and supported me with his broad knowledge while also giving me the freedom to develop my own ideas. I would also like to thank my co-examiners Prof. Hermann Matthies and Prof. Jan Hesthaven who kindly agreed to be the readers of my thesis.

I owe special thanks to my colleagues Bastian Pentenrieder and Claude Gittelson, with whom I shared my office during the last two years and had many fruitful and encouraging discussions (mathematical and otherwise), and to Roman Andreev, who programmed and carried out some of the numerical examples presented in this thesis. Besides my colleagues mentioned above, I would also like to thank all the other members and former members of the Seminar for Applied Mathematics. All of you made my time at ETH a wonderful and most enjoyable experience.

But most of all I would like to thank my parents for their constant support, encouragement and trust throughout the past years. Without them none of this would be possible.

The present thesis was partially supported by the Swiss National Science Foundation under grant No. 200021-120290 which is gratefully acknowledged.

Marcel Bieri
Abstract

We consider a stochastic Galerkin and collocation discretization scheme for solving elliptic PDEs with random coefficients and forcing term, which are assumed to depend on a finite, but possibly large number of random variables.

Both methods consist of a hierarchic wavelet discretization in space and a sequence of hierarchic approximations to the law of the random solution in probability space. In the Galerkin setting, the stochastic approximations are conducted by a best-$N$-term polynomial chaos approximation while in the collocation setting we use interpolation operators based on a Smolyak grid of Gauss points. In both approaches, we cover the case of bounded random variables as well as unbounded Gaussian random variables as input parameters. In a sparse tensor product fashion, we then compose the levels of spatial and stochastic approximations, resulting in a substantial reduction of overall degrees of freedom.

Numerical analysis is then used to estimate the convergence rates of the sparse tensor stochastic Galerkin and collocation methods, depending on the regularity of the random inputs. Numerical examples illustrate the theoretical results and indicate superiority of this novel sparse tensor product approximation compared to the ‘full tensor’ approaches used so far and the Monte Carlo method.
Zusammenfassung

Wir betrachten ein stochastisches Galerkin- und Kollokationsverfahren für die Lösung partieller Differentialgleichungen mit zufälligen Koeffizienten und Kräften, von welchen man annimmt, dass sie von einer möglicherweise grossen aber endlichen Anzahl von Zufallsvariablen abhängen.


Mittels numerischer Analysis erhalten wir Abschätzungen für die Konvergenzraten im Galerkin- und Kollokationsverfahren. Numerische Experimente illustrieren die theoretischen Resultate und lassen die Überlegenheit dieser neuartigen düntensorprodukt-Approximationen im Vergleich zu den 'volltensorprodukt'-Approximationen und der Monte-Carlo Methode erkennen.
**Contents**

Introduction xiii

1 Preliminaries 1
   1.1 Product probability spaces .......................... 1
   1.2 Tensor products of Hilbert spaces ................... 2
   1.3 Bochner spaces .................................... 3

2 Uncertainty quantification 5
   2.1 Preliminaries ..................................... 5
   2.2 Karhunen-Loève expansion ............................ 6
      2.2.1 Definition of the KL-expansion ................. 7
      2.2.2 KL eigenvalue decay ............................. 8
      2.2.3 KL eigenfunction bounds ......................... 11
      2.2.4 The lognormal case ............................. 12
   2.3 Polynomial Chaos ................................... 13
      2.3.1 Hermite polynomials of Gaussian random variables .... 14
      2.3.2 Wiener chaos expansions ......................... 14
      2.3.3 Generalized polynomial chaos ................... 15

3 Problem formulation 17
   3.1 Problem setting and notation ....................... 17
   3.2 Assumptions on the problem ........................... 18
      3.2.1 Well-posedness .................................. 18
      3.2.2 Finite-dimensional noise ........................ 18
      3.2.3 Independence .................................... 19
      3.2.4 Growth at infinity ................................ 20
      3.2.5 Stochastic regularity ............................ 20
   3.3 Model problem ...................................... 21
      3.3.1 Well-posedness .................................. 21
      3.3.2 Continuous dependence on input data ............. 24
      3.3.3 Growth at infinity ................................ 26
      3.3.4 Stochastic regularity ............................ 27
Introduction

Many engineering models of physical phenomena are subject to significant data uncertainties. We mention subsurface flow, soil mechanics, earthquake engineering, to name but a few. These uncertainties are usually crudely categorized into aleatory and epistemic uncertainties, see e.g. [13]. Aleatory uncertainties are understood as inherent variabilities of the system data parameters due to unpredictable effects, such as atmospheric conditions or subsurface properties of an aquifer in the study of groundwater flows. Epistemic uncertainties, on the other hand, are understood as model uncertainties, which are due to a fundamental lack of knowledge of the processes and quantities identified with the system. Neglecting epistemic uncertainties, we will exclusively consider PDEs with inherent parameter uncertainties, which are then often modeled as random fields [3, 61], resulting in stochastic partial differential equations.

The goal of our computations will be the approximation of statistical quantities, such as mean value and correlation, of the modeled process. Numerical solution strategies for PDEs with random field inputs follow three major steps. First, the random input fields are approximately parametrized by a finite, but possibly large number of random variables. This can e.g. be achieved by expanding the random fields into a Karhunen-Loève expansion which is eventually truncated. Then, a numerical scheme to solve the resulting high-dimensional deterministic PDEs is used to approximate the solution, depending now on space and time variables as well as on the set of input parameters. Finally, the solution is reconstructed as a random field by some form of post-processing and the statistical quantities of interest are computed. In this work, we will primarily focus on the second step, hence on the development of efficient strategies to solve PDEs depending on a large set of alterable input parameters. These algorithms can be grouped into two broad classes.

First, so-called nonintrusive schemes: here, existing deterministic solvers of the PDE of interest are used without any modification as a building block in an outer loop, where some form of sampling of the random parameter space is used to generate a set of particular input realizations to be processed by the deterministic PDE solver, leading to corresponding outputs of the random solution from which the desired statistics are recovered. Here, we find the Monte Carlo (MC) sampling strategies, stochastic collocation, as well as certain high-order polynomial chaos (PC) methods, which are based on spectral representations of the random fields’ in- and outputs. In the Monte Carlo method [21],

\[\text{Monte Carlo (MC) sampling strategies, stochastic collocation, as well as certain high-order polynomial chaos (PC) methods, which are based on spectral representations of the random fields’ in- and outputs.}\]
a large set of i.i.d. parameter samples is generated, based on their prescribed statistics, and the solution statistics derived from the set of computed solutions of the (deterministic) PDE obtained from inserting these data samples. Due to the generally slow convergence of the Monte Carlo method, the collocation approach has recently attracted a lot of attention. Introduced independently in [4] and [67], the stochastic collocation method, unlike MC, doesn’t choose the samples randomly, but in a deterministic way, based on the random inputs’ probability density functions. To overcome the curse of dimension imposed by the possibly large numbers of input parameters, the work [67] already proposed the use of Smolyak grids to reduce the number of collocation points. This was further analyzed and developed in [43] and [42]. A different collocation approach was proposed in [9], using an ANOVA based selection of collocation points to handle the high dimensionality. Here, the set of stochastic input parameters is divided into (overlapping) groups of much smaller cardinality, resulting in an efficient approximation of the solution’s random behavior.

Second, so called intrusive schemes: here, approximants to the law of the random solutions are intertwined with existing deterministic solvers at an earlier stage of the algorithm. Popular representatives of this class of algorithms are stochastic Galerkin and perturbation methods. In stochastic Galerkin, the solution is projected onto pairings of spatial and stochastic discretization spaces. A fundamental mathematical framework for the stochastic Galerkin has been laid in [5], using finite elements in space and polynomial chaos in the random parameter domain as discretizations, thus often also referred to as (generalized) polynomial chaos method (gPC), e.g. [69, 68]. It already substantiated mathematically the potential superiority of the stochastic Galerkin approach over MC type methods. Multi-element polynomial chaos methods have then been considered in [62, 63], which can be seen as an $h$-version of the gPC approach, where the probability domain is partitioned into smaller cells. Perturbation approaches lead to deterministic, but high-dimensional equations for the approximation of the $k$-th moment of the modeled process, hence aiming directly at computing the statistical moments. Using sparse tensor finite element techniques, these equations can be solved efficiently in time and memory, see [54, 55].

All of the above algorithms, except the perturbation method, consist of a sequence of stochastic approximations, e.g. polynomial chaos or collocation interpolation operators, to the law of the random solution and a spatial approximation, e.g. by finite elements. They exhibit an overall complexity, i.e. total number of degrees of freedom, of $O(N_D \times N_\Omega)$, where $N_D$ denotes the number of degrees of freedom of the spatial discretization and $N_\Omega$ the number of stochastic degrees of freedom. This is very prohibitive, especially if a fine resolution of the spatial behavior is required, e.g. due to short correlation lengths in the input random fields. The main idea of this work is to choose suitable hierarchic approximations in space and random parameter domain and combine them in a sparse tensor product fashion, leading to algorithms of $O(N_D \log N_\Omega + N_\Omega \log N_D)$.
Introduction

overall complexity, and hence a considerable reduction in computation time and memory requirement. We will use numerical analysis to tailor the stochastic approximations to the levels of the deterministic discretization for the stochastic Galerkin and stochastic collocation approach and provide convergence rates for both schemes. Our analysis will cover inputs of bounded random variables as well as Gaussian ones, which are often used in practical applications.

The outline of the thesis is as follows. In Chapter 1, we will provide some preliminary constructions and results, which will repeatedly be used throughout the rest of the work. The parametrization of random fields will then be the aim of Chapter 2. Here, we will in particular discuss the Karhunen-Loève expansion of a random field and give a short introduction into polynomial chaos. In Chapter 3, we will introduce the class of problems under consideration and impose some necessary assumptions on it, on which the numerical analysis in the following chapters will rely. An example of a problem meeting these criteria will be provided. In Chapter 4, we will formulate the sparse tensor stochastic Galerkin and sparse tensor stochastic collocation method and present a first result on the overall complexity of these algorithms. The construction of hierarchic wavelet bases for the spatial approximation is then provided in Chapter 5. The construction will explicitly be carried out for piecewise linear wavelets in one and two dimensions. In Chapter 6 we will then analyze the sparse stochastic collocation method. We will provide a hierarchic sequence of interpolation operators, based on a Smolyak grid of Gauss points, and discuss their adaption to the wavelet levels introduced in the previous chapter. In a similar manner, we will then discuss the wavelet levels introduced in the previous chapter. In a similar manner, we will then discuss the sparse stochastic Galerkin method in chapter 7. Here, we will present a hierarchic stochastic approximation by polynomial chaos in the spirit of a best-N-term approximation and prove algebraic convergence rates. Finally, in Chapter 8, we will discuss issues regarding the implementation of these algorithms and give numerical examples, which confirm the theoretical results of the previous chapters.
1 Preliminaries

We will first briefly review some of the important constructions and notations used throughout the present work. In particular countable products of probability spaces to describe the random input data and the concept of Bochner spaces. Bochner spaces and countable tensor products of probability spaces will turn out to be the natural function spaces of solutions to the stochastic PDEs under consideration.

1.1 Product probability spaces

In this section, we will review the theory of products of probability spaces. The material presented here follows closely the one provided in §9 of [6] and will therefore not further be referenced.

We denote by \((\Omega, \Sigma, P)\) a probability space with \(\Omega\) denoting the outcomes, \(\Sigma\) the sigma-algebra of possible events and \(P\) a probability measure on \(\Sigma\), hence satisfying \(P(\Omega) = 1\).

Assume we have a sequence of probability spaces \((\Omega_n, \Sigma_n, P_n)\) where \(n \in \mathbb{N}\). For a subset \(J \subset \mathbb{N}\) define \(\Omega_J := \bigtimes_{n \in J} \Omega_n\) as the Cartesian product of the \(\Omega_n\)'s with \(n \in J\). In particular, if \(J = \mathbb{N}\) we write \(\Omega := \Omega_J\). If \(J = \{j\}\) consists only of one single element, we write \(\Omega_j := \Omega_{\{j\}}\). We define the projection operator \(p_J : \Omega \to \Omega_J\) as the restriction of \(\omega \in \Omega\) to \(\Omega_J\). The product

\[
\Sigma := \bigotimes_{n \in \mathbb{N}} \Sigma_n
\]

of sigma-algebras is then defined as the smallest sigma-algebra such that any of the projections \(p_j\) is \(\Sigma\)-\(\Sigma_j\) measurable. The product measure \(P\) is defined as the unique measure on \(\bigotimes_{n \in \mathbb{N}} \Sigma_n\) s.t. for any finite set \(J \subset \mathbb{N}\) and arbitrary events \(E_j \in \Sigma_j (j \in J)\) it holds that

\[
P \left( p_J^{-1} \left( \bigotimes_{j \in J} E_j \right) \right) = \prod_{j \in J} P_j(E_j).
\]
P is then called the product measure of \( \{P_n\}_{n \in \mathbb{N}} \) and denoted by \( \bigotimes_{n \in \mathbb{N}} P_n \). The probability space

\[
\bigotimes_{n \in \mathbb{N}} (\Omega_n, \Sigma_n, P_n) := \left( \bigotimes_{n \in \mathbb{N}} \Omega_n, \bigotimes_{n \in \mathbb{N}} \Sigma_n, \bigotimes_{n \in \mathbb{N}} P_n \right)
\]

is then called the product probability space of the triplets \( \{(\Omega_n, \Sigma_n, P_n)\}_{n \in \mathbb{N}} \).

**Remark 1.1.1.** We note here, that the above definition extends also to uncountable products of probability spaces, where \( J \subset I \) is then a subset of some uncountable index set \( I \).

### 1.2 Tensor products of Hilbert spaces

This section introduces the tensor product between two Hilbert spaces, following the construction given in [48, Ch. II.4].

Let \((H_1, < \cdot, \cdot >_{H_1})\) and \((H_2, < \cdot, \cdot >_{H_2})\) be two Hilbert spaces with associated inner products. For each \( \varphi_1 \in H_1, \varphi_2 \in H_2 \), let \( \varphi_1 \otimes \varphi_2 \) denote the bilinear form, acting on \( H_1 \times H_2 \), by

\[
(\varphi_1 \otimes \varphi_2) < \psi_1, \psi_2 > := < \psi_1, \varphi_1 >_{H_1} < \psi_2, \varphi_2 >_{H_2}.
\]

Let \( \mathcal{E} \) then be the set of all finite linear combinations of such forms, and define an inner product on \( \mathcal{E} \) by

\[
< \varphi_1 \otimes \varphi_2, \psi_1 \otimes \psi_2 >_{\mathcal{E}} := < \varphi_1, \psi_1 >_{H_1} < \varphi_2, \psi_2 >_{H_2}
\]

and extending it through linearity to \( \mathcal{E} \). It can be shown, that \( < \cdot, \cdot >_{\mathcal{E}} \) is well defined and positive definite. The tensor product of \( H_1 \) and \( H_2 \), denoted by

\[
H_1 \otimes H_2,
\]

is then defined as the completion of \( \mathcal{E} \) w.r.t. the inner product \( < \cdot, \cdot >_{\mathcal{E}} \). It can be shown, that if \( \{\varphi_k\}_{k \in \mathbb{N}} \) and \( \{\psi_l\}_{l \in \mathbb{N}} \) are orthonormal bases for \( H_1 \) and \( H_2 \) respectively, then \( \{\varphi_k \otimes \psi_l\}_{k,l \in \mathbb{N}} \) is an orthonormal basis for \( H_1 \otimes H_2 \).

**Remark 1.2.1.** The construction given above immediately extends to the tensor product between \( n \) Hilbert spaces

\[
\bigotimes_{k=1}^{n} H_k = H_1 \otimes H_2 \otimes \cdots \otimes H_n.
\]
1.3 Bochner spaces

Bochner spaces are a generalization of $L^p$-spaces to functions taking values in a Banach space $X$ rather than in $\mathbb{R}$. To define the Bochner spaces, however, we first need to extend the notions of measurability and integrability to Banach-valued functions, see e.g. [10, 20, 71].

Let $(T, \Sigma, \mu)$ be a measure space with $\Sigma$ being a sigma-algebra over $T$ and $\mu$ a measure on $\Sigma$. Let $X$ be a Banach space with associated norm $\| \cdot \|_X$. A function $s : T \to X$ is called simple, if it has the form

$$s(t) = \sum_{i=1}^{n} \chi_{E_i}(t)u_i, \quad t \in T,$$

where each $E_i$ is a $\mu$-measurable subset of $T$ and $u_i \in X$ for $i = 1, \ldots, n$. A function $f : T \to X$ is called (strongly) measurable, if there exist simple functions $s_k : T \to X$, s.t.

$$s_k(t) \to f(t) \quad \text{for } \mu - \text{a.e. } t \in T$$

If $s(t) = \sum_{i=1}^{n} \chi_{E_i}(t)u_i$ is a simple function, we define

$$\int_T s(t) d\mu(t) := \sum_{i=1}^{n} \mu(E_i)u_i.$$

We further say the (strongly) measurable function $f : T \to X$ is summable, if there exists a sequence $\{s_k\}_{k=1}^{\infty}$ of simple functions, s.t.

$$\int_T \|s_k(t) - f(t)\|_X d\mu(t) \to 0 \quad \text{as } k \to \infty.$$

In that case, we define the Bochner integral

$$\int_T f(t) d\mu(t) = \lim_{k \to \infty} \int_T s_k(t) d\mu(t)$$

of $X$-valued functions $f$ over $T$. The following result will be used later on when we prove convergence rates of our algorithms.

**Theorem 1.3.1** (Bochner). A strongly measurable function $f : T \to X$ is summable if and only if $t \to \|f(t)\|_X$ is summable. In this case it holds

$$\left\| \int_T f(t) d\mu(t) \right\|_X \leq \int_T \|f(t)\|_X d\mu(t).$$
For a proof we refer to [71, Ch. V.5].

Given $1 \leq p \leq \infty$ we now define the Bochner spaces

$$L^p_\mu(T; X) = \left\{ f : T \rightarrow X : \int_T \| f(t) \|^p_X \, d\mu(t) < \infty \right\}$$  \hspace{1cm} (1.1)

with norm $\| f \|_{L^p_\mu(T; X)} := \left( \int_T \| f \|^p_X \, d\mu(t) \right)^{1/p}$ in the case of $1 \leq p < \infty$ and

$$L^\infty_\mu(T; X) = \left\{ f : T \rightarrow X : \text{ess sup}_{t \in T} \| f(t) \|_X < \infty \right\}$$  \hspace{1cm} (1.2)

with norm $\| f \|_{L^\infty_\mu(T; X)} := \text{ess sup}_{t \in T} \| f(t) \|_X$ if $p = \infty$, where the essential supremum is taken w.r.t. $\mu$. In the case where $p = 2$ and $X$ is a separable Hilbert space, we have that

$$L^2_\mu(T; X) \simeq L^2_\mu(T) \otimes X,$$  \hspace{1cm} (1.3)

where $\otimes$ denotes the tensor product between Hilbert spaces, defined in Section 1.2. In particular, $L^2_\mu(T; X)$ is itself again a Hilbert space, see e.g. [36, Ch. 1].
2 Uncertainty quantification

The computational quantitative characterization and reduction of uncertainties in a physical model, called uncertainty quantification (UQ), is a key ingredient in developing numerical schemes to describe the random behavior of the process being modeled mathematically up to a given accuracy. As explained in the introduction, we distinguish model uncertainties, i.e. incomplete knowledge about the physical process, and parameter uncertainties, i.e. incomplete knowledge about the system parameters. Mathematical methods for quantifying model uncertainties include e.g. fuzzy set theory, e.g. [72, 74, 40], and possibility theory, e.g. [73, 19]. Here, we stay within Kolmogorov’s mathematical formalism of probability. Parameter uncertainties, which are the focus of this thesis, are then very often modeled as random fields [3, 61]. However, spectral methods, such as polynomial chaos (PC) [65] and Karhunen-Loève (KL) [38] expansions of random fields have recently gained more and more attraction. In the following, we will briefly review the theory of KL- and PC-expansions and provide results, which will be used in the subsequent chapters.

2.1 Preliminaries

For the mathematical modeling of uncertainty in the model parameters, denote by $(\Omega, \Sigma, P)$ a complete probability space with $\Omega$ denoting the outcomes, $\Sigma$ the sigma-algebra of possible events and $P$ a probability measure. Furthermore, let $D \subset \mathbb{R}^d$ for $d = 1, 2, \ldots$ be a bounded, physical domain. In the following denote by

$$a(\omega, \mathbf{x}) : \Omega \times D \rightarrow \mathbb{R}$$

a random field, i.e. a jointly measurable function from $D \times \Omega$ to $\mathbb{R}$ w.r.t. the Borel sigma-algebra in $D$ and $\mathbb{R}$ and $\Sigma$ in probability space.

Assumption 2.1.1. For the random field $a$, the mean field

$$E_a(\mathbf{x}) = \int_{\Omega} a(\omega, \mathbf{x}) dP(\omega) \quad (2.1)$$

5
and covariance
\[ V_a(x, x') = \int_{\Omega} (a(\omega, x) - \mathbb{E}_a(x))(a(\omega, x') - \mathbb{E}_a(x')) \, dP(\omega) \quad (2.2) \]
are known.

An equivalent assumption would be that the mean field \( \mathbb{E}_a \) and the 2-point-correlation
\[ C_a(x, x') = \int_{\Omega} a(\omega, x)a(\omega, x') \, dP(\omega) \quad (2.3) \]
are known, since
\[ V_a(x, x') = C_a(x, x') - \mathbb{E}_a(x)\mathbb{E}_a(x'). \]

Note that for the mean and covariance to exist, we must require that \( a \) has finite second moments, i.e. \( a \in L^2_p(\Omega; L^2(D)) \). We will now define what we will later refer to as an admissible covariance function.

**Definition 2.1.2.** A covariance function \( V_a(x, x') \in L^2(D \times D) \), given by (2.2), is said to be admissible, if it is symmetric and positive definite in the sense that for any \( n \in \mathbb{N}_0 \leq n \sum_{k=1}^{n} \sum_{j=1}^{n} c_k V_a(x_k, x_j) \bar{c}_j \forall x_k, x_j \in D, c_k, c_j \in \mathbb{C}. \)

This property will be used later on to ensure that the Carleman operator associated to \( V_a \) defined in (2.5) ahead has real positive eigenvalues. Many covariance functions appearing in practice, such as Gaussian or exponential ones, are admissible. For an introduction into the theory of positive definite functions it is referred to [49, 50] where also many examples of widely used covariances are given.

### 2.2 Karhunen-Loève expansion

The Karhunen-Loève expansion can be understood as a Fourier representation of a random field, in which the spatial and stochastic parts are naturally separated into an infinite number of random variables \( \alpha_i(\omega) \) and functions \( N_i : D \subset \mathbb{R}^d \rightarrow \mathbb{R} \) (for example finite element shape functions), in the sense of
\[ a(\omega, x) = \sum_{m \geq 0} N_m(x)\alpha_m(\omega). \]

Obviously, there exist infinitely many such representations, see e.g. [34]. The Karhunen-Loève expansion, however, turns out to be an optimal approximation of the random field \( a \) in the mean square sense if truncated after the first, say \( M \), terms, see e.g. [27, 38] and also Section 2.2.2 ahead. It is therefore our preferred choice.
2.2 Karhunen-Loève expansion

2.2.1 Definition of the KL-expansion

The covariance operator of a random field $a \in L^2_p(\Omega; L^2(D))$ is

$$V_a : L^2(D) \rightarrow L^2(D), \quad (V_a u)(x) := \int_D V_a(x, x') u(x') \, dx'. \quad (2.5)$$

Given an admissible covariance function $V_a(x, x')$ in the sense of Definition 2.1.2, the associated covariance operator $V_a$ is a symmetric, non-negative and compact integral operator. It therefore has a countable sequence of eigenpairs $(\lambda_m, \varphi_m)_{m \geq 1}$

$$V_a \varphi_m = \lambda_m \varphi_m, \quad m = 1, 2, ... \quad (2.6)$$

where the sequence of real and positive KL-eigenvalues $\lambda_m$ is enumerated with decreasing magnitude and is either finite or tends to zero as $m \rightarrow \infty$, i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ (with multiplicities counted). The KL-eigenfunctions $\varphi_m(x)$ are assumed to be scaled, s.t.

$$\int_D \varphi_m(x) \varphi_n(x) \, dx = \delta_{mn}, \quad m, n = 1, 2, ... \quad (2.7)$$

i.e. they are $L^2(D)$-orthonormal. For special covariances, e.g. exponential or triangular ones, and on simple domains, the eigenpairs are known explicitly [27]. For more general covariances or more complicated domains $D$, the computation of the KL eigenpairs (2.6) can e.g. efficiently be carried out by means of a (generalized) Fast Multipole Method, see [56].

**Definition 2.2.1** (Karhunen-Loève expansion). The Karhunen-Loève (KL) expansion of a random field $a(x, \omega)$ with finite mean (2.1) and covariance (2.3), which is admissible in the sense of Definition 2.1.2, is given by

$$a(\omega, x) = \mathbb{E}_a(x) + \sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x) Y_m(\omega). \quad (2.8)$$

The family of random variables $(Y_m)_{m \geq 1}$ is determined by

$$Y_m(\omega) = \frac{1}{\sqrt{\lambda_m}} \int_D (a(\omega, x) - \mathbb{E}_a(x)) \varphi_m(x) \, dx. \quad (2.9)$$

Once the $\varphi_m(x)$ are available, probability densities $\rho_m$ of the $Y_m$ in (2.9) may be estimated from sample input fields $a(\omega, x)$ — the KL-eigenfunctions are, in effect, a tool to process such sampling data via (2.9).

One immediately verifies that

$$\mathbb{E}[Y_m] = 0, \quad \mathbb{E}[Y_m Y_n] = \delta_{mn}, \quad \forall m, n \geq 1, \quad (2.10)$$
i.e. the $Y_m$’s are centered with unit variance and pairwise uncorrelated. In the case where the random variables are Gaussian, (2.10) implies that the $Y_m$’s are independent.

In order to be able to numerically handle the KL-expansion, the series is truncated after $M$ terms, and we define

$$a_M(\omega, x) := \mathbb{E}_a(x) + \sum_{m=1}^{M} \sqrt{\lambda_m} \varphi_m(x) Y_m(\omega).$$

(2.11)

### 2.2.2 KL eigenvalue decay

The (truncated) KL-series (2.11) converges in $L^2(\Omega; L^2(D))$ to $a$ as shown in [38] and is an optimal approximation of the random field $a$ in the mean square sense, that is for any other linear combination $\tilde{a}_M$ of $M$ functions, the error $\|a - \tilde{a}_M\|_{L^2(\Omega; L^2(D))}$ is not smaller than for the KL-expansion, see e.g. [27]. Due to (2.7) and (2.10), the $L^2(\Omega; L^2(D))$ error of truncating the KL expansion is given by

$$||a - a_M||_{L^2(\Omega; L^2(D))} = \mathbb{E} \left[ \int_D (a(\omega, x) - a_M(\omega, x))^2 \, dx \right] = \sum_{m>M} \lambda_m.$$ 

(2.12)

If the sequence $(\sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)})_{m \geq 1}$ is summable and $(Y_m)_{m \geq 1}$ is uniformly bounded in $L^\infty(\Omega)$, then the convergence can easily be shown to be even uniform in $L^2(\Omega \times D)$. Precisely, we have the pointwise estimate

$$||a - a_M||_{L^\infty(\Omega \times D)} \leq C \sum_{m>M} ||\varphi_m||_{L^\infty(D)} \sqrt{\lambda_m},$$

(2.13)

where the constant $C > 0$ is independent of $M$. Estimates on the KL eigenvalue decay and $L^\infty(D)$ bounds on the eigenfunctions are therefore crucial to obtain a good a-priori control over the error of truncating the KL-expansion after $M$ terms. These bounds are in turn highly dependent on the regularity of the covariance (2.2). First results on the eigenvalue decay of integral operators with positive definite kernels in $D \subset \mathbb{R}^1$ have been proved by J.B. Reade [46, 47, 37, 35]. The results presented in the following are an extension, and sometimes also sharpened version, of those results to the case $d > 1$. Parts of it can also be found in [56].

**Definition 2.2.2.** A covariance function $V_a : D \times D \to \mathbb{R}$ is said to be piecewise analytic/smooth/H$p,q$ on $D \times D$ ($p, q \in [0, \infty]$), if there exists a partition $\mathcal{D} = \{D_j\}_{j=1}^J$ of $D$ into a finite sequence of simplices $D_j$ and a finite family $\mathcal{G} = \{G_j\}_{j=1}^J$ of open sets in $\mathbb{R}^d$ such that

$$\mathcal{D} = \bigcup_{j=1}^J D_j, \quad D_j \subset G_j \quad \forall 1 \leq j \leq J.$$
2.2 Karhunen-Loève expansion

and such that \( V_a|_{D_j \times D_j'} \) has an extension to \( G_j \times G_{j'} \) which is analytic in \( G_j \times G_{j'} \) and smooth in \( G_j \times G_{j'} \) is in \( H^{p,q}(G_j \times G_{j'}) := (L^p(G_j) \otimes L^q(D)) \cap (L^2(D) \otimes H^q(G_{j'})) \) for any pair \((j, j')\).

Proposition 2.2.3. Let \( V_a \in L^2(D \times D) \) be an admissible covariance in the sense of Definition (2.1.2).

1. If \( V_a \) is piecewise \( H^{t,t}(D \times D) \) in the sense of Definition 2.2.2, then for the eigenvalues of (2.6) it holds
   \[
   \lambda_m \leq C m^{-t/d}, \quad m \geq 1,
   \]
   with a constant \( C > 0 \) independent of \( m \).

2. If \( V_a \) is piecewise analytic in the sense of Definition 2.2.2, then the eigenvalues of (2.6) admit the bound
   \[
   \lambda_m \leq C e^{-c_1 m^{1/d}}, \quad m \geq 1,
   \]
   with \( C > 0 \) independent of \( m \).

3. If \( V_a \) is a Gaussian covariance, i.e.
   \[
   V_a(x, x') = \sigma^2 e^{-\frac{|x-x'|^2}{\gamma^2 \text{diam}(D)^2}}, \quad (x, x') \in D \times D,
   \]
   with \( \sigma, \gamma > 0 \) referring to the standard deviation and correlation length, respectively, then the eigenvalues satisfy
   \[
   0 < \lambda_m \leq C \frac{\sigma^2 (1/\gamma)^{m^{1/d}}}{\gamma^2 \Gamma\left(\frac{1}{2} m^{1/d}\right)}, \quad \forall \ m \geq 1,
   \]
   where \( \Gamma(\cdot) \) denotes the Gamma function [1, Ch. 6] and \( C > 0 \) is independent of \( M \).

The case of Gaussian covariances (2.16) is particularly interesting, since, as a consequence of the central limit theorem, it is considered in many applications. The fact that this function can be extended to an entire function in \( C^d \) leads to the faster than exponential eigenvalue decay (2.17).

Corollary 2.2.4. If \( V_a \) is piecewise smooth in the sense of Definition 2.2.2, then, due to (2.14), the eigenvalues decay at least as
   \[
   \lambda_m \leq C m^{-s}, \quad m \geq 1
   \]
   for any given \( s > 0 \), with \( C > 0 \) independent of \( m \).
Proof of Proposition 2.2.3. The first two assertions (2.14) and (2.15) have already been proved in [56, Prop. 2.18 and 2.21]. Although also the Gaussian eigenvalue decay (2.17) has already been stated in [56], a proof is missing there and therefore given here for completeness. Let $H$ be a Hilbert space on $D$ and denote by $B(H)$ the set of bounded linear operators in $H$. In [56, Lemma 2.16] it has been shown that for a symmetric, non-negative and compact operator $C \in B(H)$ with eigenpair sequence $(\lambda_m, \varphi_m)_{m \geq 1}$, it holds

$$
\lambda_{m+1} \leq \|C - C_m\|_{B(H)},
$$

where $C_m \in B(H)$ denotes an operator of rank at most $m$. We define

$$
C = \sigma^2 e^{\frac{|x - y|^2}{\gamma \text{diam}(D)^2}} = \sigma^2 \sum_{k=0}^{\infty} \frac{1}{k!} \frac{1}{(-\gamma \text{diam}(D))^{2k}} |x - y|^{2k}
$$

and

$$
C_{\bar{m}} = \sigma^2 e^{\frac{|x - y|^2}{\gamma \text{diam}(D)^2}} = \sigma^2 \sum_{k=0}^{\bar{m}} \frac{1}{k!} \frac{1}{(-\gamma \text{diam}(D))^{2k}} |x - y|^{2k}.
$$

(2.19)

Denote by $C$ and $C_{\bar{m}}$ the Carleman operators which are, by means of (2.5), associated to $C$ and $C_{\bar{m}}$, respectively. It follows that

$$
\|C - C_{\bar{m}}\|_{B(H)} \lesssim \|C - C_{\bar{m}}\|_{L^\infty(D \times D)}
$$

$$
\lesssim \sigma^2 \sum_{k=\bar{m}+1}^{\infty} \frac{1}{k!} \frac{1}{(\gamma \text{diam}(D))^{2k}} \text{diam}(D)^{2k}
$$

$$
\lesssim \frac{1}{(\bar{m} + 1)!} \gamma^{2(\bar{m} + 1)}
$$

$$
\lesssim \frac{1}{(\bar{m} + 1)!} \gamma^{2(\bar{m} + 1)}. \quad \text{(2.19)}
$$

The rank of the Carleman operator defined by (2.19) can be estimated as

$$
\text{rank}(C_{\bar{m}}) \leq \text{dim span}\{x_1^{\alpha_1} \cdots x_d^{\alpha_d} : \alpha_1 + \cdots + \alpha_d \leq 2\bar{m}\} = \binom{2\bar{m} + d}{d}.
$$

Hence, for a given $m \in \mathbb{N}$ we choose $\bar{m}$ such that

$$
\binom{2\bar{m} + d}{d} \leq m \leq \binom{2(\bar{m} + 1) + d}{d},
$$

from which it follows that $\bar{m} \sim \frac{1}{2} m^{1/d}$ as $m \to \infty$. This completes the proof of Proposition 2.2.3. \qed
2.2.3 KL eigenfunction bounds

The first result states that the regularity of the covariance \( V_a \) implies the corresponding regularity of the KL eigenfunctions and can be found in [56, Prop. 2.23].

**Proposition 2.2.5.** Assume that the covariance \( V_a \) is piecewise analytic/ smooth/\( H^{p,q} \) w.r.t. \( D \) in the sense of Definition 2.2.2. Then the Karhunen-Loève eigenfunctions are analytic/smooth/\( H^p \) in every \( D_j \in D \).

As motivated at the beginning of the previous paragraph, \( L^\infty(D) \) bounds on the KL eigenfunctions \( \varphi_m \) are crucial to obtain a pointwise control over the error stemming from truncating the KL expansion. We will first discuss the case of covariances \( V_a \) with finite Sobolev regularity and then state the corresponding result for smooth/analytic/Gaussian covariances as a corollary.

**Proposition 2.2.6.** Assume that the covariance \( V_a \) is piecewise \( H^{t,t}(D \times D) \) w.r.t. \( D \) with \( t > d/2 \). Then \( \varphi_m \in H^{t}((D_j^2)) \) for every \( D_j \in D \) and for every \( \varepsilon \in (0, t - d/2] \) there exists a constant \( C > 0 \), depending on \( \varepsilon, d \) but not on \( m \), such that

\[
\|\varphi_m\|_{L^\infty(D)} \leq C \lambda_m^{-\varepsilon} m \geq 1
\] (2.20)

**Corollary 2.2.7.** Assume that the covariance \( V_a \) is piecewise smooth w.r.t. \( D \). Then, for every \( s > 0 \) there exists a constant \( C > 0 \) depending on \( s, d \), but not on \( m \), such that

\[
\|\varphi_m\|_{L^\infty(D)} \leq C \lambda_m^{-s} m \geq 1
\] (2.21)

**Proof of Proposition 2.2.6.** The \( H^t(D) \) regularity of the eigenfunctions is due to Proposition 2.2.5. The Sobolev embedding theorem [2] for all \( t^* \in (d/2, t] \) then yields

\[
\|\varphi_m\|_{C^{0}(\overline{D})} \leq C \|\varphi_m\|_{H^{t^*}(D)} \quad \forall m \geq 1,
\] (2.22)

with a positive constant \( C \) depending on \( t^*, d \). It remains to bound the norm on the right hand side of the inequality. It is well-known that fractional order Sobolev spaces \( H^{t^*}(D) \) can be obtained by complex interpolation [60] between \( L^2(D) \) and \( H^t(D) \) for any \( t > t^* \) and we therefore have the Riesz-Thorin-type inequality [60, §1.9.3]

\[
\|u\|_{H^{t^*}(D)} \leq \|u\|_{H^t(D)}^{t^*/t} \|u\|_{L^2(D)}^{1-t^*/t}
\]

for any function \( u \in H^t(D) \). Choosing now \( u = \varphi_m \) we have, due to the \( L^2 \)-normalization that

\[
\|\varphi_m\|_{H^{t^*}(D)} \leq \|\varphi_m\|_{H^t(D)},
\] (2.23)
From the eigenvalue equation
\[ \varphi_m(x) = \frac{1}{\lambda_m} \int_D C_a(x,x')\varphi_m(x') \, dx', \]
it follows by differentiating and using the Cauchy-Schwartz inequality
\[ \|\partial^\alpha \varphi_m\|_{L^\infty(D)} \leq C_\alpha \lambda_m^{-1}. \tag{2.24} \]
Combining now (2.24), (2.23) and (2.22) yields
\[ \|\varphi_m\|_{L^\infty(D)} \leq C \lambda_m^{-t^*/t}, \]
which implies (2.20). \(\square\)

Combining now Proposition 2.2.3 with Proposition 2.2.6, allows us to get a pointwise control of the error stemming from truncating the KL-expansion (2.13). Precisely,

**Corollary 2.2.8.** Assume that the covariance \(V_a\) is piecewise \(H^{t,t}(D \times D)\) w.r.t. \(D\) with \(t > d/2\).
\[ \sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)} \leq C m^{-s} \tag{2.25} \]
where
\[ s = \frac{1}{2d}(t - d - 2\varepsilon) \tag{2.26} \]
for an arbitrary \(\varepsilon \in (0, t - d/2]\) and where the constant \(C\) in (2.25) is independent of \(m\).

### 2.2.4 The lognormal case

Lognormal random fields are random fields of the form \(a(\omega, x) = e^{g(\omega, x)}\), with \(g\) being Gaussian. Therefore, its logarithm can be expanded in a Karhunen-Loève series as before:
\[ \ln a(\omega, x) = \mathbb{E}_a^{\log}(x) + \sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x) Y_m(\omega) \tag{2.27} \]
with \(Y_m \sim \mathcal{N}(0, 1)\). Similarly to (2.11), we then define the truncated exponential KL series as
\[ a_M(\omega, x) := e^{\mathbb{E}_a^{\log}(x) + \sum_{m = 1}^M \sqrt{\lambda_m} \varphi_m(x) Y_m(\omega)}. \tag{2.28} \]
Since \(a - a_M = a_M(\frac{a}{a_M} - 1)\), we obtain for the \(L^2\) truncation error
\[ \|a - a_M\|_{L^2([\Omega;L^2(D))} = \|a_M\|_{L^2([\Omega;L^2(D))} \|e^{\sum_{m > M} \sqrt{\lambda_m} \varphi_m Y_m} - 1\|_{L^2([\Omega;L^2(D))}. \]
Since the random variables \(Y_m\) are Gaussian, they can take values on the whole real axis. Therefore, to obtain a pointwise estimate on the truncation error, we have to control
2.3 Polynomial Chaos

their behavior towards \(-\infty\) and \(\infty\). To this end, we introduce the Gaussian measure, denoted by

\[
\mathcal{G}_\sigma := \bigotimes_{m=1}^{\infty} G_{m,\sigma_m},
\]

(2.29)

where \(\sigma = (\sigma_1, \sigma_2, \ldots)\) and \(G_{m,\sigma_m}\) is univariate Gaussian measure with variance \(\sigma_m^2\) and zero mean imposed on \(Y_m\), see Section 1.1 for the definition of product measures.

Remark 2.2.9. It can be proved that the countable product measure in (2.29) defines a Gaussian measure on \(\mathbb{R}^{\infty} = \times_{m=1}^{\infty} \mathbb{R}\) and moreover that this is essentially the unique Gaussian measure with prescribed mean and variance, see [45, Sec. 1.5].

Denote by \(\ell_{\text{exp}}^1\) the space of all sequences \(\sigma \in \mathbb{R}^N\), with \(\sigma_m = e^{s_m}\) where \(\{s_m\}_{m \in \mathbb{N}}\) is a summable sequence of positive real numbers. If \(\sigma \in \ell_{\text{exp}}^1\), we have

\[
L^2_{\mathcal{G}_\sigma}(\Omega) \subset L^2_{\mathcal{G}_1}(\Omega)
\]

with \(\|u\|_{L^2_{\mathcal{G}_1}(\Omega)} \leq \|u\|_{L^2_{\mathcal{G}_\sigma}(\Omega)}\). Then it can be shown, using a result from [33] that \(\mathcal{G}_\sigma\) is equivalent to the standard Gaussian measure \(\mathcal{G}_1\) on \(\mathbb{R}^{\infty}\) with zero mean and unit variance, see [28, Prop. 2.11]. We denote by

\[
\frac{d\mathcal{G}_\sigma}{d\mathcal{G}_1} = \prod_{m=1}^{\infty} \frac{1}{\sigma_m} e^{-\frac{1}{2} \sum_{m=1}^{\infty} (\sigma_m^2 - 1)y_m^2}
\]

(2.30)

the respective Radon-Nikodym derivative. This allows us to define weighted \(L^\infty\) spaces as

\[
L^\infty_{\mathcal{G}_\sigma}(\Omega \times D) := \{f : \Omega \times D \rightarrow \mathbb{R} : \text{ess sup}_{\Omega \times D} \left| f \left( \frac{d\mathcal{G}_\sigma}{d\mathcal{G}_1} \right)^{-1} \right| < \infty \}
\]

(2.31)

with \(\|f\|_{L^\infty_{\mathcal{G}_\sigma}} := \text{ess sup}_{\Omega \times D} \left| f \left( \frac{d\mathcal{G}_\sigma}{d\mathcal{G}_1} \right)^{-1} \right|\). The pointwise KL truncation error can then be written as

\[
\|a - a_M\|_{L^\infty_{\mathcal{G}_\sigma}(\Omega \times D)} = \|a_M\|_{L^\infty_{\mathcal{G}_\sigma}(\Omega \times D)} \|e^{\sum_{m>M} \sqrt{\lambda_m}Y_m} - 1\|_{L^\infty_{\mathcal{G}_\sigma}(\Omega \times D)}.
\]

(2.32)

2.3 Polynomial Chaos

We saw in the previous section that a prerequisite to write a square-summable random field into a Karhunen-Loève expansion, is the knowledge of its mean and covariance. This can clearly be assumed for the stochastic input parameters but not for the response
statistics of the physical system. Hence, an alternative expansion is needed to describe the solution’s random behavior. One such alternative is polynomial chaos (PC), first introduced by N. Wiener [65] for Gaussian random processes. The idea of PC is to take polynomial combinations of known random variables (e.g. the KL random variables of the inputs) as a basis for the expansion. Its deterministic coefficients have then to be found by minimizing some norm of the error resulting from a finite representation, e.g. by Galerkin projection.

2.3.1 Hermite polynomials of Gaussian random variables

Assume \( \{Y_m(\omega)\}_{m=1}^{\infty} \) is a set of Gaussian random variables satisfying (2.10) and \( P \) is the standard Gaussian measure \( G_1 \). Denote by \( \mathbb{N}_c^N \) the space of all sequences \( \nu = (\nu_1, \nu_2, \ldots) \) of non-negative natural numbers with compact support, i.e. where the set \( \text{supp}(\nu) = \{m \in \mathbb{N} : \nu_m \neq 0\} \) is finite. For \( \nu \in \mathbb{N}_c^N \), define by

\[
\mathcal{H}_\nu(Y) = e^{\frac{1}{2}Y^\top Y} (-1)^{\nu_0} \frac{\partial^{|\nu|}}{\partial Y_1^{\nu_1} \partial Y_2^{\nu_2} \cdots} e^{-\frac{1}{2}Y^\top Y}
\]

the multivariate Hermite polynomial, where \( Y = (Y_1, Y_2, \ldots) \). Note that this definition is meaningful, since only a finite number of \( \nu_m \)'s are different from zero. It is proved in [14, Thm. 9.1.5] that the system \( \{\mathcal{H}_\nu\}_{\nu \in \mathbb{N}_c^N} \) is orthogonal and complete in \( L^2_P(\Omega) \), i.e. every function \( f \in L^2_P(\Omega) \) can be written as a sum

\[
f = \sum_{\nu \in \mathbb{N}_c^N} f_\nu \mathcal{H}_\nu, \quad f_\nu \in \mathbb{R}
\]

converging in \( L^2_P(\Omega) \).

2.3.2 Wiener chaos expansions

Denote by \( \hat{\pi}_p \) the space spanned by multivariate Hermite polynomials in \( \{Y_m(\omega)\}_{m=1}^{\infty} \) with total degree at most \( p \), i.e.

\[
\hat{\pi}_p = \left\{ f = \sum_{\nu \in \mathbb{N}_c^N : |\nu| \leq p} f_\nu \mathcal{H}_\nu : f_\nu \in \mathbb{R} \right\}.
\]

Furthermore, let \( \pi_p \) then represent the set of all polynomials in \( \hat{\pi}_p \), \( L^2_P \)-orthogonal to \( \hat{\pi}_{p-1} \), called polynomial chaos of order \( p \). The space \( \pi_p = \text{span}\{\pi_p\} \subset L^2_P(\Omega) \) is called
2.3 Polynomial Chaos

the $p$-th homogeneous chaos and is given by

$$
\pi_p = \left\{ f = \sum_{\nu \in \mathbb{N}^{|\nu|} \atop |\nu|=p} f_{\nu} \mathcal{H}_{\nu} : f_{\nu} \in \mathbb{R} \right\}
$$

Consequently, $\pi_p$ is often also called Hermite chaos. It follows, [14, Thm. 9.1.7] that

$$
L^2_P(\Omega) = \bigoplus_{p \geq 0} \tilde{\pi}_p,
$$

which is called the Wiener-Itô decomposition of $L^2_P(\Omega)$. Therefore, any square integrable Gaussian random variable $f \in L^2_P(\Omega)$, can be represented as

$$
f = \sum_{n=0}^{\infty} \sum_{\nu \in \mathbb{N}^{|\nu|} \atop |\nu|=n} f_{\nu} \mathcal{H}_{\nu}, \quad f_{\nu} \in \mathbb{R}
$$

with convergence in $L^2_P(\Omega)$. Truncating the expansion, such that only polynomials with total order less than $p$ are taken into account, gives rise to the so-called polynomial chaos expansion of order $p$ of $f$, i.e.

$$
\tilde{f} = \sum_{n=0}^{p} \sum_{\nu \in \mathbb{N}^{|\nu|} \atop |\nu|=n} f_{\nu} \mathcal{H}_{\nu}, \quad f_{\nu} \in \mathbb{R}
$$

2.3.3 Generalized polynomial chaos

Although N. Wiener originally proposed the polynomial chaos in terms of Gaussian random variables, the same idea also carries over to non-Gaussian ones which is then often referred to generalized polynomial chaos (gPC), introduced in [69, 51]. Given a set of i.i.d. random variables $\{Y_m(\omega)\}_{m=1}^{\infty}$ with product measure $P = \otimes_{m=1}^{\infty} P_m$, the spaces $\hat{\pi}_p$, $\bar{\pi}_p$ are then constructed by means of the respective (multivariate) orthogonal polynomials, e.g. Legendre polynomials

$$
\frac{1}{2^n n!} \frac{\partial^n}{\partial Y_{i_1} \cdots \partial Y_{i_p}} \prod_{j=1}^{p} Y_{i_j} \prod_{j=1}^{p} (Y_{i_j}^2 - 1)
$$

in case of uniform random variables $Y_m$, leading to so-called Legendre chaos, and likewise for other probability distributions. In the course of the discretization discussed in Chapter 7, we shall in particular use Legendre and Hermite expansions. We emphasize, however, that also for other probability densities, other polynomial systems can be used in exactly the same fashion, see e.g. [51] for more details.
2 Uncertainty quantification
3 Problem formulation

In the present chapter, we will introduce the class of elliptic problems under consideration and the notation used throughout the remainder of the thesis. We follow [4] and sharpen and generalize some of the results therein. The concepts of Bochner spaces and tensor products of Hilbert spaces, introduced in Chapter 1, appear as natural function spaces in the given framework. We will then impose several assumptions on the problem, on which the algorithms developed in Chapters 4–7 are relying. The chapter is then concluded by providing an example of a problem satisfying all of the assumptions which will then, in a slightly simplified form, also serve as a model problem for the following chapters to illustrate the numerical methods developed in this work.

3.1 Problem setting and notation

The problem under consideration is a stochastic elliptic boundary value problem of the form

\[
\begin{align*}
\mathcal{L}(u) &= f & \text{in } D, \\
\mathcal{B}(u) &= g & \text{on } \partial D
\end{align*}
\]

(3.1)

where \( \mathcal{L} \) is an elliptic differential operator, depending on one or more random parameters, e.g. a random diffusion coefficient \( a(\omega, x) \), and \( \mathcal{B} \) is a (deterministic) boundary operator. The forcing term \( f(\omega, x) \) can also assumed to be random, whereas, on the other hand, we do not consider randomness of the Dirichlet data \( g \) or the domain \( D \), as e.g. done in [29, 30, 70] and references therein.

In the following, we will often interpret the random coefficients \( a, f \) and the solution \( u \) not as real-valued, measurable functions on \( \Omega \times D \), but as random variables, taking values in a suitable Banach space \( \mathcal{W}(D) \) of solutions, if we attempt to solve (3.1) for given realizations \( a(\omega, \cdot) \) and \( f(\omega, \cdot) \) of the stochastic inputs. Hence, we introduce the Bochner spaces

\[
L^2_P(\Omega; \mathcal{W}(D)) \quad \text{and} \quad L^\infty_P(\Omega; \mathcal{W}(D)),
\]

(3.2)

as defined in Section 1.3.
Note that if $W(D)$ is a separable Hilbert space, then so is $L^2_P(\Omega; W(D))$ and, by virtue of (1.3), there exists an isomorphism, s.t.

$$L^2_P(\Omega; W(D)) \cong L^2_P(\Omega) \otimes W(D)$$

(3.3)

where $\otimes$ denotes the tensor product between separable Hilbert spaces, as introduced in Section 1.2.

### 3.2 Assumptions on the problem

The numerical methods presented in this work rely on a number of assumptions imposed on the problem (3.1), which will be introduced next. Some of these assumptions are mandatory while others may be relaxed. At some points below, we will give remarks on possible generalizations and also refer to other works where these have been addressed.

#### 3.2.1 Well-posedness

The first assumption ensures that the problem is well posed in the sense of unique solvability.

**Assumption 3.2.1.** The random coefficients $a$ and the forcing term $f$ are such that the existence and uniqueness of a solution $u(\omega, x) \in L^2_P(\Omega; W(D))$ to (3.1) is guaranteed.

#### 3.2.2 Finite-dimensional noise

The next assumption restricts the model problem (3.1) to ones where the stochastic behavior of the input parameters can be described by a finite-dimensional random vector $(Y_1, \ldots, Y_M)$.

**Assumption 3.2.2.** The stochastic parameters $a$ and $f$ depend only on a finite number $M$ of random variables $Y_m : \Omega \to \mathbb{R}$, i.e.

$$a(\omega, x) = a(Y_1, \ldots, Y_M, x) \quad \text{and} \quad f(\omega, x) = f(Y_1, \ldots, Y_M, x).$$

(3.4)

This may seem very restricting at a first glance, since random fields are usually only properly described by an infinite number of random variables. However, since our goal is an approximation to the statistical moments of the solution, we only need to describe the random field to a sufficiently high accuracy, see also Section 3.3.2 ahead, which is usually possible by only taking a finite number of random variables into account, as we have seen in (2.12) and (2.13).
Remark 3.2.3. Assumption 3.2.2 can be omitted for the Galerkin method presented in Chapter 7. In fact, the Galerkin method presented there turns out to be dimension-adaptive, meaning that it automatically selects the relevant, finitely many input variables to achieve the desired target accuracy.

### 3.2.3 Independence

**Assumption 3.2.4.**

i) The family \((Y_m)_{m \geq 1} : \Omega \to \mathbb{R}\) is independent,

ii) with each \(Y_m(\omega)\) is associated a probability space \((\Omega_m, \Sigma_m, P_m), m \in \mathbb{N}\), with the following properties:

a) the probability measure \(P_m\) admits a probability density function

\[
\rho_m : \text{Ran}(Y_m) := \Gamma_m \longrightarrow [0, \infty),
\]

such that \(dP_m(\omega) = \rho_m(y_m) dy_m, \ m \in \mathbb{N}, y_m \in \Gamma_m\) and

b) the sigma algebras \(\Sigma_m\) are subsets of the Borel sets of the interval \(\Gamma_m\), i.e. \(\Sigma_m \subseteq \Sigma(\Gamma_m)\).

Due to the independence we can interpret the \(Y_m\)'s as different coordinates in probability space and hence parametrize \(\Omega\) by the vector

\[
y = (y_1, y_2, \ldots) \in \Gamma := \Gamma_1 \times \Gamma_2 \times \cdots
\]

rather than \((Y_1(\omega), Y_2(\omega), \ldots)\). The vector \(y\) is then equipped with the product probability density \(\rho = \prod_{m \geq 1} \rho_m : \Gamma \to \mathbb{R}\). In the following we will therefore write \(a(y, x)\) instead of \(a(\omega, x)\) and likewise for \(u\) and \(f\). We will then identify

\[
L^2_\rho(\Gamma; \mathcal{W}(D)) = L^2_\rho(\Omega; \mathcal{W}(D)) \quad \text{and} \quad L^\infty_\rho(\Gamma; \mathcal{W}(D)) = L^\infty_\rho(\Omega; \mathcal{W}(D)). \quad (3.5)
\]

**Remark 3.2.5.** The random variables \(Y_m\) will in general not be independent, except for Gaussian random fields (2.16). One possible remedy has been proposed in [4], by introducing auxiliary probability density functions \(\tilde{\rho}_m\) with \(\|\rho_m/\tilde{\rho}_m\|_{L^\infty(\Gamma)} < \infty\), such that the random variables \(Y_m\) are independent with respect to \(\tilde{\rho} = \prod_{m \geq 1} \tilde{\rho}_m\). An alternative way to treat the case of dependent random variables has been proposed in [57], where an orthonormal system of polynomials are constructed, which then serve as a basis for the PC representation of the response process.
3 Problem formulation

3.2.4 Growth at infinity

In the previous paragraph, the parameter domains $\Gamma_m = \text{Ran}(Y_m)$ have been introduced. These may either be bounded, e.g. in the case of uniform random variables, or unbounded, which includes in particular the case of Gaussian or exponential random variables $Y_m$. In the case of unbounded parameter domains, however, the growth at infinity has to be controlled. We define the function space

$$C_0^\chi(\Gamma; W(D)) := \left\{ v : \Gamma \rightarrow W(D) : v \text{ cont. in } y, \sup_{y \in \Gamma} \| \chi(y)v(y) \|_{W(D)} < \infty \right\}, \quad (3.6)$$

where

$$\chi(y) = \prod_{m=1}^M \sigma_m(y_m) \leq 1 \quad \text{and} \quad \chi_m(y_m) = \begin{cases} 1 & \text{if } \Gamma_m \text{ bounded} \\ e^{-\alpha_m|y_m|} & \text{for some } \alpha_m > 0 \text{ if } \Gamma_m \text{ unbounded} \end{cases} \quad (3.7)$$

The associated norm is then given by $\|v\|_{C_0^\chi} := \sup_{y \in \Gamma} \| \chi(y)v(y) \|_{W(D)}$.

**Assumption 3.2.6.**

i) $u \in C_0^\chi(\Gamma; W(D))$ and

ii) the joint probability density $\rho$ satisfies

$$\rho(y) \leq C_\rho e^{-\sum_{m=1}^M (\delta_m y_m)^2}, \quad \forall y \in \Gamma \quad (3.8)$$

for a constant $C_\rho > 0$ and $\delta_m$ strictly positive if $\Gamma_m$ is unbounded and zero otherwise.

**Remark 3.2.7.** The decay parameter $\chi$ (3.7) and the density $\rho$ (3.8) are chosen in such a way that

$$C_0^\chi(\Gamma; W(D)) \subset L^2_\rho(\Gamma; W(D)).$$

Therefore, for the inclusion still to hold, other choices of $\chi$ and $\rho$ could be considered as well. Relaxing the parameter $\chi$ would then impose stronger assumptions on $\rho$ and vice versa.

3.2.5 Stochastic regularity

Finally we need to make some regularity assumption on the stochastic behavior of the solution. To simplify the notation we write

$$y_m^* := (y_1, \ldots, y_{m-1}, y_{m+1}, \ldots) \in \Gamma_m^*, \quad \text{where} \quad \Gamma_m^* := \prod_{j \neq m} \Gamma_j.$$
Assumption 3.2.8. For each \( m \in \{1, \ldots, M\} \) there exists \( \tau_m > 0 \), such that the solution \( u(y_m, y_m^*, x) \), as a function of \( y_m \in \Gamma_m \), admits an analytic extension to the closed region of the complex plane
\[
\Sigma(\Gamma_m, \tau_m) := \{ z \in \mathbb{C} : \text{dist}(z, \Gamma_m) \leq \tau_m \}.
\] (3.9)

3.3 Model problem

To conclude this section, we will present one example where the above assumptions hold true. This example, in a simplified form, will serve as our model problem throughout the rest of the paper.

Given probability spaces \((\Omega_i, \Sigma_i, P_i), i = 1, 2, 3\) and random fields
\[
a(\omega, x) : \Omega_1 \times D \to \mathbb{R}, \quad c(\omega, x) : \Omega_2 \times D \to \mathbb{R}, \quad f(\omega, x) : \Omega_3 \times D \to \mathbb{R},
\]
we consider the following stochastic diffusion-reaction problem
\[
\begin{cases}
-\text{div}(a(\omega, x)\nabla u(\omega, x)) + c(\omega, x)u(\omega, x) = f(\omega, x) & D, \quad P-a.e. \omega \in \Omega, \\
u(\omega, x)|_{x \in \partial D} = 0,
\end{cases}
\] (3.10)
where \((\Omega, \Sigma, P) := \bigotimes_{i=1}^3 (\Omega_i, \Sigma_i, P_i)\) denotes the product probability space as defined in Section 1.1. Here, \( a \) and \( c \) denote the diffusivity and reaction rate, respectively and \( f \) a stochastic source term. In accordance with Assumption 3.2.4, we assume that the random coefficients are independent. We choose \( W(D) = H^1_0(D) \) and by multiplying with a test function and integrating by parts, we obtain the variational formulation:

Find \( u \in L^2_P(\Omega; H^1_0(D)) \) s.t. \( \forall v \in L^2_P(\Omega; H^1_0(D)) \) it holds
\[
b(u, v) = l(v),
\] (3.11)
with
\[
b(u, v) = \mathbb{E} \left[ \int_D a \nabla u \nabla v + cuv \, dx \right], \quad l(v) = \mathbb{E} \left[ \int_D fv \, dx \right],
\] (3.12)
where we suppress the dependence of the coefficients and functions on \((\omega, x) \in \Omega \times D\) for notational convenience.

3.3.1 Well-posedness

Assumption 3.2.1 is satisfied if for the coefficients \( a, c, f \) it holds
3 Problem formulation

i) \( a \in L^2_p(\Omega; L^2(D)) \) is positive and bounded away from zero almost surely, i.e. there exists \( a^* > 0 \) such that

\[
P \left\{ \omega \in \Omega : a^* \leq \text{ess inf}_{x \in D} a(\omega, x) \right\} = 1.
\]

(3.13)

ii) \( c \in L^2_p(\Omega; L^2(D)) \) is non-negative almost surely, i.e.

\[
P \left\{ \omega \in \Omega : 0 \leq \text{ess inf}_{x \in D} c(\omega, x) \right\} = 1.
\]

(3.14)

iii) \( f \) is square-integrable with respect to \( P \), i.e.

\[
E[\|f\|_{L^2(D)}^2] < \infty.
\]

(3.15)

Indeed, by introducing the Hilbert space

\[
\mathcal{H}_{a,c} := \{ v : \Gamma \rightarrow H^1_0(D) : E \left[ \int_D a|\nabla v|^2 + cv^2 \, dx \right] < \infty \}
\]

with the norm \( \|v\|_{\mathcal{H}_{a,c}}^2 := E[\int_D a|\nabla v|^2 + cv^2 \, dx] \), from an application of the Lax-Milgram Lemma it follows the existence and uniqueness of a solution \( u \in \mathcal{H}_{a,c} \) to (3.11). Precisely, consider the bilinear form \( b(u, v) \) given in (3.12). It follows immediately

\[
b(u, v) \leq \|u\|_{\mathcal{H}_{a,c}} \|v\|_{\mathcal{H}_{a,c}} \quad \text{and} \quad b(u, u) \geq \|u\|_{\mathcal{H}_{a,c}}^2,
\]

i.e. \( b(\cdot, \cdot) \) is continuous and coercive with continuity and coercivity constant equal to one. It remains to show that \( l(\cdot) \) (3.12) is continuous w.r.t. \( \mathcal{H}_{a,c} \):

\[
l(v) = E \left[ \int_D f(\omega, x)v(\omega, x) \, dx \right]
\leq E \left[ \|f(\omega, \cdot)\|_{L^2(D)} \|v(\omega, \cdot)\|_{L^2(D)} \right]
\leq \frac{C_P}{\sqrt{a^*}} E \left[ \|f(\omega, \cdot)\|_{L^2(D)} \|\sqrt{a(\omega, \cdot)}\nabla v(\omega, \cdot)\|_{L^2(D)} \right]
\leq \frac{C_P}{\sqrt{a^*}} \|f\|_{L^2_p(\Omega; L^2(D))} \|v\|_{\mathcal{H}_{a,c}},
\]

where \( C_P \) denotes the Poincaré constant, i.e. \( \|v\|_{L^2(D)} \leq C_P \|\nabla v\|_{L^2(D)} \), and the norm \( \|f\|_{L^2_p(\Omega; L^2(D))} \) is finite due to (3.15). Hence, by Lax-Milgram, the existence of a unique solution \( u \in \mathcal{H}_{a,c} \) to (3.11) is guaranteed \( P \)-almost surely. Furthermore, since

\[
\|u\|_{L^2_p(\Omega; H^1_0(D))} \leq \frac{C_P}{\sqrt{a^*}} \|u\|_{\mathcal{H}_{a,c}},
\]

(3.17)
the space $\mathcal{H}_{a,c}$ is continuously embedded in $L^2_P(\Omega; H^1_0(D))$, i.e.

$$\mathcal{H}_{a,c} \hookrightarrow L^2_P(\Omega; H^1_0(D)) \quad (3.18)$$

and it follows the existence of a solution $u \in L^2_P(\Omega; H^1_0(D))$ to (3.11).

**Remark 3.3.1.** As pointed out in [4], it is possible to relax the condition (3.13) above to a lower bound which is itself again a random variable, i.e. there exists $a_*(\omega)$ s.t.

$$P\{\omega \in \Omega : 0 < a_*(\omega) < \text{ess inf}_{x \in D} a(\omega, x)\} = 1. \quad (3.19)$$

This case is of particular interest, since it covers the case of lognormal random coefficients $a$ (2.27). However, the fact that $a_*(\omega)$ can take values arbitrary close to zero, calls for stronger regularity assumptions. In the following, denote by $P$ always the standard $M$-fold Gaussian measure $G_1$, as constructed in Section 2.2.4. Given $\sigma \in L^\exp_1$, consider the following variational problem w.r.t. the measure $G_\sigma$: find $u \in L^2_P(\Omega; H^1_0(D))$, s.t.

$$b^\sigma(u,v) = l^\sigma(v), \quad \forall v \in L^2_P(\Omega; H^1_0(D)), \quad (3.20)$$

where $b^\sigma$ and $l^\sigma$ are defined as

$$b^\sigma_M(u_M, v) = \mathbb{E}^\sigma \left[ \int_D a \nabla u_M \cdot \nabla v + cuv \, dx \right]$$

and

$$l^\sigma(v) = \mathbb{E}^\sigma \left[ \int_D f v \, dx \right]$$

where we again suppressed the dependence on $(\omega, x)$ for reasons of readability and where $\mathbb{E}^\sigma$ denotes the expectation taken w.r.t. $G_\sigma$. Accordingly, define $\mathcal{H}^\sigma_{a,c}$ as in (3.16) w.r.t. $G_\sigma$. The well-posedness and hence the existence of a solution in $\mathcal{H}^\sigma_{a,c}$ can then be shown in essentially the same way as above, except that we need stronger regularity assumptions on $f$. We have

$$l^\sigma(v) = \mathbb{E}^\sigma \left[ \int_D f(\omega, x)v(\omega, x) \, dx \right] \leq \mathbb{E}^\sigma \left[ \|f(\omega, \cdot)\|_{L^2(D)}\|v(\omega, \cdot)\|_{L^2(D)} \right] \leq C_P \mathbb{E}^\sigma \left[ \sqrt{\|a(\omega, \cdot)\|_{L^2(D)}} \|\nabla v(\omega, \cdot)\|_{L^2(D)} \right]$$

and by twice using Hölder’s inequality

$$\leq C_P \|f\|_{L^2_{G_\sigma}(\Omega, L^2(D))}^{1/2p} \|v\|_{L^2_{G_\sigma}(\Omega)}^{1/2p} \|\mathcal{H}^\sigma_{a,c}\| \quad (3.21)$$
where $1/p + 1/q = 1$ and $p,q \geq 1$. Hence, if $1/\sqrt{\sigma} \in L_{G_a}^{2q}(\Omega)$ for some $q \geq 1$, then we must require $f \in L_{G_a}^{2q}(\Omega; L^{2p}(D))$ for $p = (1 - 1/q)^{-1}$. If we assume that $a$ can be expanded in a lognormal KL expansion (2.27), then a lower bound to $a$ is given by

$$a_\ast(\omega) := e^{\min \mathbb{L}^{\log}(x) - \sum_{m \geq 1} \sqrt{\mathcal{N}\psi_m(x)} |Y_m(\omega)|}.$$  

(3.22)

We then have

$$\|u\|_{L_p^2(\Omega; H_0^1(D))} \leq \left|\frac{1}{a_\ast}\right|_{L_\infty^1(\Omega \times D)} \|u\|_{\mathcal{H}_{a,c}},$$

(3.23)

where $P$ denotes the standard Gaussian measure and with $L_{G_a}^\infty(\Omega \times D)$ as defined in (2.31), which implies

$$\mathcal{H}_{a,c}^\ast \hookrightarrow L_p^2(\Omega; H_0^1(D)),$$

i.e. $\mathcal{H}_{a,c}$ is continuously embedded in $L_p^2(\Omega; H_0^1(D))$, where $\hookrightarrow$ denotes the inclusion map.

### 3.3.2 Continuous dependence on input data

A main requirement for the stable numerical approximation of (3.10) is continuous dependence of the solution $u$ on the random input data. Moreover, this also justifies the approximation of the random fields $a, c, f$ by a truncated KL-expansion. We will abbreviate the notions of the norms as follows:

$$\| \cdot \|_{L_p^r(\mathcal{W})} = \| \cdot \|_{L_p^r(\mathcal{W}(D))} \quad \text{and} \quad \| \cdot \|_{L_\infty} = \| \cdot \|_{L_\infty(\Omega \times D)}$$

(3.24)

unless otherwise indicated.

**Proposition 3.3.2.**

i) Denote by $a, c, f$ and $\tilde{a}, \tilde{c}, \tilde{f}$ two sets of bounded random fields satisfying (3.13)- (3.15) and by $u$ and $\tilde{u}$ the respective unique solution to (3.10). Then it holds

$$\|u - \tilde{u}\|_{L_p^2(H_0^1)} \lesssim (\|a - \tilde{a}\|_{L_\infty} + \|c - \tilde{c}\|_{L_\infty}) \|f\|_{L_p^2(L^2)} + \|f - \tilde{f}\|_{L_p^2(L^2)},$$

(3.25)

where the constant depends only on $a_\ast$ and $\text{diam}(D)$.

ii) Denote by $a, c, f$ and $\tilde{a}, \tilde{c}, \tilde{f}$ two sets of lognormal random fields (2.27) satisfying (3.19), (3.14) and by $u$ and $\tilde{u}$ the respective unique solution to (3.10). By $G_a$ denote a Gaussian measure as constructed in Section 2.2.4 and let $P$ be the standard Gaussian measure $G_1$. It holds

$$\|u - \tilde{u}\|_{L_p^2(H_0^1)} \lesssim \left(\|a - \tilde{a}\|_{L_{G_a}^{2q}(L^{2q})}^{1/2q} + \|c - \tilde{c}\|_{L_{G_a}^{2q}(L^{2q})}^{1/2q}\right) \|f\|_{L_{G_a}^{2p'}(L^{2p'})}^{1/2q}$$

$$+ \|f - \tilde{f}\|_{L_{G_a}^{2p'}(L^{2p'})}^{1/2q}$$

(3.26)
3.3 Model problem

with \( p', q > 1 \) satisfying \( \frac{1}{p'} + \frac{1}{q} < 1 \) and the constant depending only on \( a_*, q, p' \) and \( \text{diam}(D) \).

**Proof.**

i) First, we treat the case of bounded random fields. Denote by \( b(\cdot, \cdot) \) and \( \tilde{b}(\cdot, \cdot) \), the bilinear forms (3.12) w.r.t. \( a, c \) and \( \tilde{a}, \tilde{c} \), respectively, and by \( l(\cdot) \), \( \tilde{l}(\cdot) \) the linear forms (3.12) w.r.t. \( f \) and \( \tilde{f} \), respectively. Hence we have the variational problem of finding \( u, \tilde{u} \in L^2_P(\Omega; H^1_0(D)) \), respectively, s.t.

\[
\begin{align*}
    b(u, v) &= l(v) \\
    \tilde{b}(\tilde{u}, v) &= \tilde{l}(v)
\end{align*}
\]

for all test functions \( v \in L^2_P(\Omega; H^1_0(D)) \). Due to (3.13) it holds

\[
\begin{align*}
    \|u - \tilde{u}\|_{L^2_P(H^1_0)}^2 &\leq \frac{1}{a_*} b(u - \tilde{u}, u - \tilde{u}) \\
    &= b(u, u - \tilde{u}) - b(\tilde{u}, u - \tilde{u}) + \tilde{b}(\tilde{u}, u - \tilde{u}) \\
    &= \tilde{b}(\tilde{u}, u - \tilde{u}) - b(\tilde{u}, u - \tilde{u}) + l(u - \tilde{u}) - \tilde{l}(u - \tilde{u}).
\end{align*}
\]

(3.27)

For the linear functionals we obtain

\[
\begin{align*}
    l(u - \tilde{u}) - \tilde{l}(u - \tilde{u}) &= \int_D (f - \tilde{f})(u - \tilde{u}) d\mathbf{x} P(\omega) \\
    &\leq C_P \|f - \tilde{f}\|_{L^2_P(L^2)} \|u - \tilde{u}\|_{L^2_P(H^1_0)},
\end{align*}
\]

(3.28)

where \( C_P \) denotes the Poincaré constant. Similarly, for the bilinear forms we obtain

\[
\tilde{b}(\tilde{u}, u - \tilde{u}) - b(\tilde{u}, u - \tilde{u}) \leq C_P (\|a - \tilde{a}\|_{L^\infty} + \|c - \tilde{c}\|_{L^\infty}) \|\tilde{u}\|_{L^2_P(H^1_0)} \|u - \tilde{u}\|_{L^2_P(H^1_0)}
\]

(3.29)

By observing that

\[
\|\tilde{u}\|_{L^2_P(H^1_0)} \leq \frac{1}{a_*} b(\tilde{u}, \tilde{u}) = l(\tilde{u}) \leq C_P \|f\|_{L^2_P(L^2)} \|u\|_{L^2_P(H^1_0)},
\]

(3.30)

(3.25) follows from combining (3.27)-(3.30).

ii) For the unbounded case, denote by \( b^\sigma, \tilde{b}^\sigma \) the bilinear forms from Remark 3.3.1 w.r.t. the lognormal coefficients \( a, c \) and \( \tilde{a}, \tilde{c} \), respectively and by \( l^\sigma, \tilde{l}^\sigma \) the linear forms w.r.t. \( f \) and \( \tilde{f} \), respectively. We consider the variational problem of finding \( u, \tilde{u} \in L^2_P(\Omega; H^1_0(D)) \), respectively, s.t.

\[
\begin{align*}
    b^\sigma(u, v) &= l^\sigma(v) \\
    \tilde{b}^\sigma(\tilde{u}, v) &= \tilde{l}^\sigma(v)
\end{align*}
\]

(3.29)
Problem formulation

for all test functions \( v \in L^2_p(\Omega; H^1_0(D)) \). Note that due to (3.23) it is sufficient to find a bound of \( u - \tilde{u} \) in the energy norm \( \| \cdot \|_{\mathcal{H}^2_{\alpha,c}} \) as defined in Remark 3.3.1. As in (3.27) it follows

\[
\| u - \tilde{u} \|_{\mathcal{H}^2_{\alpha,c}}^2 = \tilde{b}^\sigma(\tilde{u}, u - \tilde{u}) - b^\sigma(\tilde{u}, u - \tilde{u}) + l^\sigma(u - \tilde{u}) - \tilde{l}^\sigma(u - \tilde{u}).
\]

(3.31)

By similar arguments as in (3.21), we obtain

\[
\| u - \tilde{u} \|_{\mathcal{H}^2_{\alpha,c}}^2 \leq \| f - \tilde{f} \|_{L^2_p(\Omega; L^2_p)}^{1/2q} \left\| \frac{1}{\sqrt{a^*}} \right\|_{L^2_p(\Omega; L^2_p)}^{1/2q} \| u - \tilde{u} \|_{\mathcal{H}^2_{\alpha,c}}
\]

(3.32)

For the error in the bilinear form we have

\[
b^\sigma(\tilde{u}, u - \tilde{u}) - \tilde{b}^\sigma(\tilde{u}, u - \tilde{u}) = \int_\Omega \int_D \left( \frac{a - \tilde{a}}{\sqrt{a}} \right) \tilde{u} \left( \sqrt{a} \nabla (u - \tilde{u}) \right) d\Omega \ dG_\sigma(\omega)
\]

\[
+ \int_\Omega \int_D \left( \frac{c - \tilde{c}}{\sqrt{c}} \right) \tilde{u} \left( \sqrt{a} (u - \tilde{u}) \right) d\Omega \ dG_\sigma(\omega)
\]

and by twice using Hölder’s inequality

\[
\leq C \left\| \frac{a - \tilde{a}}{\sqrt{a}} \right\|_{L^2_p(\Omega; L^2_p)}^{1/2q} \left\| \nabla \tilde{u} \right\|_{L^2_p(\Omega; L^2_p)}^{1/2q} \| u - \tilde{u} \|_{\mathcal{H}^2_{\alpha,c}}
\]

\[
+ \left\| \frac{c - \tilde{c}}{\sqrt{c}} \right\|_{L^2_p(\Omega; L^2_p)}^{1/2q} \left\| \nabla \tilde{u} \right\|_{L^2_p(\Omega; L^2_p)}^{1/2q} \| u - \tilde{u} \|_{\mathcal{H}^2_{\alpha,c}}
\]

(3.33)

It remains to estimate the \( L^2_p(\Omega; L^2_p(D)) \)-norm of \( \nabla \tilde{u} \). To this end we use a result from [28, Thm. 3.14], stating that

\[
\| \nabla \tilde{u} \|_{L^2_p(\Omega; L^2_p(D))} \leq C_{\alpha,p,a} \| f \|_{L^2_p(\Omega; L^2_p(D))},
\]

(3.34)

provided that \( \tilde{u} \) is a solution to a diffusion equation with lognormal coefficients and \( p' > p \). Combining (3.31)-(3.34) completes the proof of (3.26).

3.3.3 Growth at infinity

As imposed by Assumption 3.2.6, in the case of unbounded random variables, we have to control the behavior of \( u \) as the parameters \( y_m \) tend towards \(-\infty \) and \( \infty \). Define the pointwise (or parametric) bilinear form

\[
B(y; u, v) = \int_D a(y, x) \nabla u(y, x) \nabla v(x) \, dx, \quad y \in \Gamma, \ u, v \in H^1_0(D)
\]

(3.35)
and linear form

\[ F(y; v) = \int_D f(y, x)v(x) \, dx, \quad y \in \Gamma, \ v \in H^1_0(D). \]  \tag{3.36}

Then the solution to the parametric deterministic problem of finding \( u(y) \in H^1_0(D) \), such that

\[ B(y; u, v) = F(y; v), \]

satisfies the pointwise estimate

\[ \|u(y)\|_{H^1_0(D)} \leq \frac{C_P}{a_+ (y)} \|f(y)\|_{L^2(D)}, \quad y \in \Gamma. \]

Hence, if we assume \( f \in C^0_\chi(\Gamma; L^2(D)) \), then, by (3.22), we clearly have \( u \in C^0_\chi(\Gamma; L^2(D)) \) with decay parameter \( \alpha_m = \tilde{\alpha}_m + \sqrt{\lambda_m} \|\varphi\|_{L^\infty(D)} \) (3.7).

### 3.3.4 Stochastic regularity

The following proposition shows that under certain assumptions on the growth of the derivatives of \( a, c, f \), the coordinate-wise analyticity as requested in Assumption 3.2.8 is ensured.

**Proposition 3.3.3.** Assume that for every \( y = (y_m, y_m^*) \) there exists \( \gamma_m < \infty \) satisfying

\[
\frac{\|\partial_{y_m} a(y)\|_{L^\infty(D)}}{a(y)} \leq \gamma_m^k k!, \quad \frac{\|\partial_{y_m} c(y)\|_{L^\infty(D)}}{c(y)} \leq \gamma_m^k k!, \quad \frac{\|\partial_{y_m} f(y)\|_{L^2(D)}}{1 + \|f\|_{L^2(D)}} \leq \gamma_m^k k! \tag{3.37}
\]

Assume further that there exists \( p, q \geq 1 \) with \( 1/p + 1/q = 1 \), such that there holds \( c \in C^0_{\chi'/p}(\Gamma; L^\infty(D)) \) and \( f \in C^0_{\chi'/q}(\Gamma; L^2(D)) \). Then the solution \( u(y_m, y_m^*; x) \) to (3.10), as a function of \( y_m, u : \Gamma_m \to C^0_{\chi_m}(\Gamma_m; H^1_0(D)) \), admits an analytic extension into the region \( \Sigma(\Gamma_m, \tau_m) \subset \mathbb{C} \) with \( 0 < \tau_m < \frac{1}{2\gamma_m} \), uniformly for all \( y_m^* \in \Gamma_m^* \).

**Proof.** We consider the problem (3.10) which can be stated in variational form as: Find \( u \in L^2_p(\Gamma; H^1_0(D)) \) s.t. \( \forall v \in H^1_0(D) \) it holds

\[ B(y; u, v) = F(y; v) \]  \tag{3.38}

with

\[ B(y; u, v) = \int_D a \nabla u \nabla v + cuv \, dx \quad F(y; v) = \int_D f v \, dx, \]  \tag{3.39}

where we suppress the dependence of the coefficients and functions on \( (y, x) \in \Gamma \times D \) to shorten the formulae below.
3 Problem formulation

We note here that Lemma 3.3.3 is a slight generalization of the result given in [4, Lemma 3.2] and hence the following proof follows closely the one given there.

Differentiating (3.38) \( k \) times with respect to \( y_m \) and using Leibniz’s rule we obtain

\[
\sum_{l=0}^{k} \binom{k}{l} \int_D \left( \partial_{y_m}^l a \nabla \partial_{y_m}^{k-l} u \nabla v + \partial_{y_m}^l c \partial_{y_m}^{k-l} uv \right) dx = \int_D \partial_{y_m}^k f v dx.
\]

For the remaining part of the proof we denote by \( \| \cdot \|_{L²} \) and \( \| \cdot \|_{L∞} \) the norms of \( L²(D) \) and \( L∞(D) \), respectively. Setting \( v = \partial_{y_m}^k u \) we obtain

\[
\left\| \sqrt{a} \nabla \partial_{y_m}^k u \right\|_{L²}^2 \leq \sum_{l=1}^{k} \binom{k}{l} \left\| \frac{\partial_{y_m}^l a}{a} \right\|_{L∞} \left\| \sqrt{a} \nabla \partial_{y_m}^{k-l} u \right\|_{L²} + \sum_{l=1}^{k} \binom{k}{l} \left\| \frac{C_p^2}{a_s} \partial_{y_m}^l c \right\|_{L∞} \left\| \partial_{y_m}^{k-l} u \right\|_{L²} + \| \partial_{y_m}^k f \|_{L²} \left\| \partial_{y_m}^k u \right\|_{L²}.
\]

Using the Poincaré inequality and dividing both sides by \( \| \sqrt{a} \nabla \partial_{y_m}^k u \|_{L²} \) gives

\[
\left\| \sqrt{a} \nabla \partial_{y_m}^k u \right\|_{L²} \leq \sum_{l=1}^{k} \binom{k}{l} \left\| \frac{\partial_{y_m}^l a}{a} \right\|_{L∞} \left\| \sqrt{a} \nabla \partial_{y_m}^{k-l} u \right\|_{L²} + \sum_{l=1}^{k} \binom{k}{l} \frac{C_p^2}{a_s} \left\| \partial_{y_m}^l c \right\|_{L∞} \left\| \partial_{y_m}^{k-l} u \right\|_{L²} + \frac{C_p}{\sqrt{a_s}} \| \partial_{y_m}^k f \|_{L²} \left\| \partial_{y_m}^k u \right\|_{L²}.
\]

Using now the assumptions (3.37) and setting \( R_k(y) = \| \sqrt{a} \nabla \partial_{y_m}^k u(y, \cdot) \|_{L²} / k! \), we obtain the following pointwise bound on \( R_k \)

\[
R_k(y) \leq \sum_{l=1}^{k} \gamma_m^l R_{k-l}(y) + \frac{C_p^2}{a_s} \| c(y, \cdot) \|_{L∞} \gamma_m^l R_{k-l}(y) + \frac{C_p}{\sqrt{a_s}} (1 + \| f(y, \cdot) \|_{L²}) \gamma_m^k.
\]

Setting \( C_1(y) := 1 + \frac{C_p^2}{a_s} \| c(y, \cdot) \|_{L∞} \) and \( C_2(y) := \frac{C_p}{\sqrt{a_s}} (1 + \| f(y, \cdot) \|_{L²}) \), we deduce by induction

\[
R_k(y) \leq C_1(y) 2 \gamma_m^k (R_0(y) + C_2(y)), \quad \forall y \in \Gamma.
\]

Since by (3.38)

\[
R_0(y) = \| \sqrt{a}(y, \cdot) \nabla u(y, \cdot) \|_{L²} \leq \frac{C_p}{\sqrt{a_s}} \| f(y, \cdot) \|_{L²} \leq C_2(y), \quad \forall y \in \Gamma,
\]

28
we finally obtain
\[
\left\| \nabla \frac{\partial^k y_m}{a_k} u(y, \cdot) \right\|_{L^2} \leq \frac{1}{\sqrt{a_s}} R_k(y) \leq \frac{2C_1(y)C_2(y)}{\sqrt{a_s}} (2\gamma_m)^k, \quad \forall y \in \Gamma. \tag{3.40}
\]

Now define the formal power series of \( u \) around \( y_m \in \Gamma_m \):
\[
u(z_m, y_m^*, x) = \sum_{k=0}^{\infty} \frac{\partial^k y_m u(y_m, y_m^*, x)}{k!} (z_m - y_m)^k.
\]
Using now (3.40), we obtain
\[
\chi_m(y_m) \| u(z_m) \|_{C^0_{\chi_m}(\Gamma_m, H^1_0)} \leq \sum_{k=0}^{\infty} \frac{\left\| \partial^k y_m u(z_m) \right\|_{C^0_{\chi_m}(\Gamma_m, H^1_0)}}{k!} \chi_m(y_m) |z_m - y_m|^k
\]
\[
\leq \frac{2 \|C_1\|_{C^0_{1/p}(\Gamma, L^\infty)} \|C_2\|_{C^0_{1/q}(\Gamma, L^2)}}{\sqrt{a_s}}
\times \sum_{k=0}^{\infty} (|z_m - y_m| 2\gamma_m)^k.
\]
Thus the formal power series converges for all \( z_m \in \mathbb{C} \) with \( \text{dist}(z_m, y_m) \leq \tau_m < 1/(2\gamma_m) \) and hence by a continuation argument in \( \Sigma(\Gamma_m, \tau_m) \) where \( \tau_m < 1/(2\gamma_m) \). By uniqueness, its limit is equal to \( u \).

The assumptions of Proposition 3.3.3 are satisfied if the random fields can be expanded into a linear (2.8) or exponential (2.27) Karhunen-Loève series. In fact, in the linear case we have that
\[
\left\| \frac{\partial^k y_m a}{a} \right\|_{L^\infty(\Gamma \times D)} \leq \begin{cases} \sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)} / a_s & k = 1, \\ 0 & k > 1, \end{cases}
\tag{3.41}
\]
such that we can choose \( \gamma_m = \sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)} / a_s \). In the lognormal case we have
\[
\left\| \frac{\partial^k y_m a}{a} \right\|_{L^\infty(\Gamma \times D)} \leq \left( \sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)} \right)^k,
\tag{3.42}
\]
and we can safely take \( \gamma_m = \sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)} \).
3 Problem formulation
4 Sparse tensor discretizations

In this chapter, sparse tensor formulations of the stochastic Galerkin and stochastic collocation methods are introduced and applied to the model problem under consideration. Here, as opposed to other works in the field of stochastic PDEs, the term ‘sparse’ refers to a sparse composition of hierarchic discretizations in both, random and spatial parameter spaces. Here and throughout the rest of the thesis, the term ‘composition’ refers to a superposition of tensor products of detail spaces in hierarchic discretizations. The chapter aims at providing a general description of the sparse tensor stochastic Galerkin and collocation method. The specification of the discretization spaces along with a thorough discussion of the resulting sparse stochastic Galerkin and collocation algorithms is then provided in Chapters 5–7. We do emphasize, however, that also other discretizations as the ones proposed there could be considered instead, e.g. $hp$-discretization in space and ANOVA-based discretizations in random parameter space [59, 9].

4.1 Model problem

To simplify the presentation in the present and following chapters, we will restrict our discussion to the stochastic diffusion problem

$$
\begin{cases}
-\text{div}(a(\omega, x)\nabla u(\omega, x)) = f(x) & \text{in } D, \\
u(\omega, x)|_{x \in \partial D} = 0,
\end{cases}
$$

(4.1)

for $P$-a.e. $\omega \in \Omega$, where the diffusion coefficient $a$ and the solution $u$ are random fields in the physical domain $D$, with $a$ satisfying (3.13). The source term $f(x) \in L^2(D)$ is assumed to be deterministic. We further assume that the random field can be expanded in either a bounded linear or exponential Karhunen-Loève series, where either of them is truncated after $M$ terms, hence, according to (2.11):

$$
a_M(\omega, x) = E_a(x) + \sum_{m=1}^{M} \psi_m(x)Y_m(\omega), \quad \psi_m(x) := \sqrt{\lambda_m}\varphi_m(x),
$$

(4.2)

or

$$
a_M(\omega, x) = e^{E_a(x)} + \sum_{m=1}^{M} \psi_m(x)Y_m(\omega), \quad \psi_m(x) := \sqrt{\lambda_m}\varphi_m(x),
$$

(4.3)
where we assume that \( \{ \| \psi_m \|_{L^\infty(D)} \}_{m \in \mathbb{N}} \) is summable.

**Remark 4.1.1.** It follows directly from Corollary 2.2.8 that the summability of the \( \psi_m \)'s in the \( L^\infty(D) \)-norm is guaranteed, if we e.g. assume that for the covariance \( V_a \) (2.2), associated with the random field \( a \) or its logarithm, respectively, it holds that \( V_a \in H^{k,d}(D \times D) \) for \( t > 3d \).

With either of the finite expansions (4.2), (4.3) in hand, we replace in (4.1) the random input \( a \) by its \( M \)-term truncated KL expansion \( a_M \) and obtain the truncated problem

\[
\begin{align*}
-\text{div}(a_M(\omega, x)\nabla u_M(\omega, x)) &= f(x) \quad \text{in} \; D, \quad P \text{- a.e.} \; \omega \in \Omega \\
\left. u_M(\omega, x) \right|_{x \in \partial D} &= 0,
\end{align*}
\]

where \( u_M \) denotes the solution to this truncated problem. The variational formulation of (4.4) with a linear coefficient (4.2) is then given by: Find \( u_M \in L^2_\rho(\Gamma; H^1_0(D)) \), s.t.

\[
b_M(u_M, v) = l(v), \quad \forall \; v \in L^2_\rho(\Gamma; H^1_0(D)),
\]

with

\[
b_M(u_M, v) = \mathbb{E} \left[ \int_D a_M(y, x)\nabla u_M(y, x) \cdot \nabla v(y, x) \, dx \right]
\]

\[
= \int_\Gamma \int_D a_M(y, x)\nabla u_M(y, x) \cdot \nabla v(y, x) \rho(y) \, dx \, dy
\]

and

\[
l(v) = \mathbb{E} \left[ \int_D f(x)v(y, x) \, dx \right] = \int_\Gamma \int_D f(x)v(y, x) \rho(y) \, dx \, dy.
\]

The unique solvability of (4.5), uniformly in the number \( M \) of terms retained in the truncated Karhunen-Loève expansion, follows from (3.13) and (2.13). More precisely, there exists \( M_0 > 0 \) and \( \gamma > 0 \) (depending on \( a_* \) and \( \sum_{m \geq 1} \| \psi_m \|_{L^\infty(D)} \), but not on \( M \)), such that

\[
\forall M \geq M_0 \quad \forall v \in L^2_\rho(\Gamma; H^1_0(D)) : \quad b_M(v, v) \geq \gamma \| v \|_{L^2_\rho(\Gamma; H^1_0(D))}^2,
\]

\[
\forall v, w \in L^2_\rho(\Gamma; H^1_0(D)) : \quad \| b_M(v, w) \| \leq \gamma^{-1} \| v \|_{L^2_\rho(\Gamma; H^1_0(D))} \| w \|_{L^2_\rho(\Gamma; H^1_0(D))}
\]

and by using Lax-Milgram, the existence of a unique solution to (4.5) follows. Note that due to the results established in Section 3.3, (4.4) satisfies all of the Assumptions 3.2.1 - 3.2.8. According to Proposition 3.3.2, we obtain for the error between the original solution \( u \) to (4.1) and the truncated solution \( u_M \) to (4.4)

\[
\| u - u_M \|_{L^2(\Gamma; H^1_0(D))} \leq C \| a - a_M \|_{L^\infty(\Gamma \times D)} \| f \|_{L^2(\Gamma; L^2(D))}.
\]
4.2 Stochastic Galerkin formulation

In the case where \( a_M \) is given by a truncated exponential KL series (4.3), we consider, as in Remark 3.3.1, the variational problem

\[
b_{M}^{\sigma}(u_M, v) = l^{\sigma}(v),
\]

(4.10)

where \( b^\sigma \) and \( l^\sigma \) are defined as in (4.6), (4.7) but with the expectation taken w.r.t. the Gaussian measure \( G_\sigma \), i.e.

\[
b_{M}^{\sigma}(u_M, v) = \int_{\Gamma} \int_{D} a_M(y, x) \nabla u_M(y, x) \cdot \nabla v(y, x) \rho_\sigma(y) \, dx \, dy
\]

(4.11)

and

\[
l^{\sigma}(v) = \int_{\Gamma} \int_{D} f(x)v(y, x)\rho(y)\rho_\sigma(y) \, dx \, dy,
\]

(4.12)

where \( \rho_\sigma \) denotes the density of \( G_\sigma \) w.r.t. the Lebesgue measure on \( \Gamma \). If \( \sigma = 1 \), we write \( \rho = \rho_1 \). Given \( \sigma, \tilde{\sigma} \in \ell^{\exp} \) with \( \tilde{\sigma} > \sigma \) component-wise, we obtain

\[
\forall M \geq M_0 \quad \forall v \in L^2_\rho(\Gamma; H^1_0(D)) : \quad |b_{M}^{\sigma}(v, v)| \geq \essinf_{y \in \Gamma} a^* \frac{d G_\sigma}{d y_1} \|v\|_{L^2_\rho(\Gamma; H^1_0(D))}^2,
\]

(4.13)

\[
\forall v, w \in L^2_{\rho_\sigma}(\Gamma; H^1_0(D)) : \quad |b_{M}^{\sigma}(v, w)| \leq \esssup_{y \in \Gamma} a^* \frac{d G_\sigma}{d y_1} \|v\|_{L^2_{\rho_\sigma}(\Gamma; H^1_0(D))} \|w\|_{L^2_{\rho_\sigma}(\Gamma; H^1_0(D))}
\]

(4.14)

Existence and uniqueness of a solution \( u \in L^2_\rho(\Gamma; H^1_0(D)) \) can then be shown as in Remark 3.3.1.

4.2 Stochastic Galerkin formulation

4.2.1 Hierarchic subspace sequences

In the stochastic Galerkin FEM (sGFEM), we discretize the variational formulation (4.5) by Galerkin projection onto a sequence of finite dimensional subspaces of

\[
L^2_\rho(\Gamma; H^1_0(D)) \approx L^2_\rho(\Gamma) \otimes H^1_0(D).
\]

(4.15)

Specifically, we choose two hierarchic families of finite dimensional subspaces

\[
V_0^\Gamma \subset V_1^\Gamma \subset \ldots \subset V_l^\Gamma \subset V_{l+1}^\Gamma \subset \ldots \subset L^2_\rho(\Gamma)
\]

(4.16)

and

\[
V_0^D \subset V_1^D \subset \ldots \subset V_{l_1}^D \subset V_{l_1+1}^D \subset \ldots \subset H^1_0(D),
\]

(4.17)
with $l_1, l_2$ being the levels of refinement. We introduce detail spaces $W^T_{l_1}$ and $W^D_{l_2}$, such that

$$W^T_0 := V^T_0 \quad \text{and} \quad V^T_{l_1} = V^T_{l_1-1} \oplus W^T_{l_1} \quad \text{for} \quad l_1 = 1, 2, \ldots$$

and

$$W^D_0 := V^D_0 \quad \text{and} \quad V^D_{l_2} = V^D_{l_2-1} \oplus W^D_{l_2} \quad \text{for} \quad l_2 = 1, 2, \ldots$$

(4.18)

where the sums are direct so that the (finite-dimensional) approximation spaces $V^T_{l_1}$ and $V^D_{l_2}$ admit a multilevel decomposition

$$V^T_{l_1} = \bigoplus_{k_1=0}^{L} W^T_{k_1} \quad \text{and} \quad V^D_{l_2} = \bigoplus_{k_2=0}^{L} W^D_{k_2}.$$  

(4.20)

In the lognormal case (4.10) we proceed analogously, except that we choose subspaces $V^T_{l_1} \subset L^p(\rho_\sigma(\Gamma))$ such that we have continuity and coercivity both w.r.t. functions in $V^T_{l_1}$, taking values in $V^D_{l_2}$, guaranteeing then for quasi-optimality of the Galerkin formulation introduced below.

The specific choice of multilevel discretizations in space and random domain will be subject of the following chapters, where we will propose a wavelet discretization in the physical domain $D$ and a polynomial chaos discretization in $L^2_\rho(\Gamma)$.

### 4.2.2 Sparse tensor stochastic Galerkin formulation

Given the subspace hierarchies $\{V^T_{l_1}\}_{l_1 \geq 0} \subset L^2_\rho(\Gamma)$ and $\{V^D_{l_2}\}_{l_2 \geq 0} \subset H^1_0(D)$ in (4.15), which admit splittings (4.18) and (4.19), we denote by

$$V^T_L \otimes V^D_L = \bigoplus_{0 \leq l_1, l_2 \leq L} W^T_{l_1} \otimes W^D_{l_2} \subset L^2_\rho(\Gamma) \otimes H^1_0(D)$$

(4.21)

the (full) tensor product space of the finite dimensional component subspaces $V^D_{l_2}$ and $V^T_{l_1}$, respectively. The (full) stochastic Galerkin formulation of (4.5) is then given by: Find

$$\bar{u}_M \in V^T_L \otimes V^D_L : \quad b_M(\bar{u}_M, v) = l(v) \quad \forall \, v \in V^T_L \otimes V^D_L.$$  

(4.22)

Denoting by $N^T_L$ and $N^D_L$ the number of stochastic and deterministic degrees of freedom, respectively, the stochastic Galerkin formulation w.r.t. $V^T_L \otimes V^D_L$ clearly uses

$$\dim (V^T_L \otimes V^D_L) = N^T_L \times N^D_L$$

(4.23)

degrees of freedom. Due to the generally large number of KL terms $M$ and therefore high dimensionality of the space $L^2_\rho(\Gamma; H^1_0(D))$, a lot of recent works were focused on
4.2 Stochastic Galerkin formulation

Figure 4.1: Illustration of the sparse tensor product space $V^T_L \otimes V^D_L$ in terms of the component detail spaces $W^T_{l_1}$ and $W^D_{l_2}$.

the reduction of $N_L^T$, by using sparse approximation techniques [22, 59, 12]. Here, we will consider these approaches and, in addition, approximate the solution to (4.5) by Galerkin projection onto sparse tensor product spaces defined by

$$V^T_L \otimes V^D_L := \bigoplus_{0 \leq l_1 + l_2 \leq L} W^T_{l_1} \otimes W^D_{l_2},$$

see also Figure 4.1 for an illustration of $V^T_L \otimes V^D_L$ in terms of the component detail spaces.

Hence, we obtain the sparse stochastic Galerkin formulation: find

$$\hat{u}_M \in V^T_L \otimes V^D_L : \quad b_M(\hat{u}_M, v) = l(v) \quad \forall v \in V^T_L \otimes V^D_L.$$
As a consequence of (4.8), (4.9), if \( M \) is sufficiently large, (4.25) defines, for every \( L \geq 0 \) a unique sGFEM approximation \( \hat{u}_M \in V^*_L \otimes V^D_L \) which, by Galerkin orthogonality, is a quasi-optimal approximation in \( L^2_\rho(\Gamma; W(D)) \) of \( u_M \) defined in (4.5):

\[
\|u_M - \hat{u}_M\|_{L^2_\rho(\Gamma; W(D))} \leq C\|u_M - \hat{v}\|_{L^2_\rho(\Gamma; W(D))} \quad \forall \hat{v} \in V^*_L \otimes V^D_L, \quad y \in \Gamma
\]  

Similarly, we obtain the sparse stochastic Galerkin formulation for the lognormal problem (4.10): find \( \hat{u}_M \in V^*_L \otimes V^D_L \):

\[
b_M^\sigma(\hat{u}_M, v) = l^\sigma(v) \quad \forall v \in V^*_L \otimes V^D_L.
\]  

Since \( V^*_L \subset L^2_\rho(\Gamma) \subset L^2_\rho(\Gamma) \), the bilinear form \( b_M^\sigma(\cdot, \cdot) \) is by virtue of (4.13), (4.14) coercive and continuous, both w.r.t. \( u, v \in V^*_L \otimes V^D_L \). As a consequence, we have quasi-optimality as in (4.26), see also [28, Thm. 4.4].

\[
\|u_M - \hat{u}_M\|_{L^2_\rho(\Gamma; W(D))} \leq C\|u_M - \hat{v}\|_{L^2_\rho(\Gamma; W(D))} \quad \forall \hat{v} \in V^*_L \otimes V^D_L.
\]  

Using the sparse tensor product space (4.24) instead of the full tensor product space (4.21), results in a reduction of the complexity (4.23) to essentially log-linear complexity. Precisely, we have

**Lemma 4.2.1.** Assume that the dimensions of the detail spaces \( W^r_{l_1} \) and \( W^D_{l_2} \) grow exponentially with respect to the levels \( l_1, l_2 \), i.e. there exist two geometric sequences \( b_{l_1}^1 \) and \( b_{l_2}^2 \), \( l_1, l_2 = 1, 2, 3, \ldots \), with bases \( b_T \) and \( b_D \) such that

\[
\dim(W^r_{l_1}) \sim b_{l_1}^1 \quad \text{and} \quad \dim(W^D_{l_2}) \sim b_{l_2}^2.
\]

Then the sparse tensor product space \( V^*_L \otimes V^D_L \) is of cardinality

\[
\dim(V^*_L \otimes V^D_L) \sim O(L^\theta \max(b_T, b_D)^{L+1}),
\]  

where \( \theta = 1 \) if \( b_T = b_D \) and zero otherwise.

**Proof.** Observe that

\[
\dim(V^*_L \otimes V^D_L) \leq c \sum_{l_1=0}^{L} \sum_{l_2=0}^{L-l_1} b_{l_1}^1 b_{l_2}^2 \leq c \sum_{l_1=0}^{L} b_{l_1}^1 b_{l_2}^2 = c \sum_{l_1=0}^{L} b_{l_1}^1.
\]  

In the case where \( b_T = b_D \), we therefore have

\[
\dim(V^*_L \otimes V^D_L) \leq c \sum_{l_1=0}^{L} b_{l_1}^{L+1} = c(L + 1)b_1^{L+1},
\]  

36
4.3 Stochastic collocation formulation

which leads to (4.29) with \( \theta = 1 \). Now we assume w.l.o.g. \( b_D > b_\Gamma \) and it follows from (4.30)

\[
\dim(V_L^\Gamma \otimes V_L^D) \leq c \sum_{l_1=0}^{L} b_1^{l_1} b_D^{-l_1+1}
\]

\[
= c b^{L+1} \sum_{l_1=0}^{L} b_1^{l_1} \left( \frac{b_\Gamma}{b_D} \right)^{l_2}
\]

\[
= c b^{L+1} \left( \frac{b_D}{b_\Gamma} \right)^{L+1} - 1
\]

Since the denominator is positive it can be absorbed in the constant and we find

\[
\dim(V_L^\Gamma \otimes V_L^D) \leq c b^{L+1}
\]

This completes the proof.

4.3 Stochastic collocation formulation

4.3.1 Hierarchic sequence of collocation operators

Recall the definition of the spatial discretization spaces \( V_L^D \) and their respective details \( W_L^D \) from (4.17) and (4.19). In addition, define the \( H^1_0 \)-projection operator \( P_L^D : H^1_0(D) \to V_L^D \) by

\[
(\nabla(u_M - P_L^D u_M), \nabla v)_{L^2} = 0 \quad \forall v \in V_L^D,
\]

where \((\cdot, \cdot)_{L^2}\) denotes the \( L^2 \) inner product on \( D \).

For each (stochastic) level \( l_1 \in \mathbb{N}_0 \), choose a discrete finite set of collocation points \( \mathcal{Y}_{l_1} \subset \Gamma \) and a set of multiindices \( \Lambda_{l_1} \subset \mathbb{N}^M \), such that the sequence is nested with increasing \( l_1 \), i.e.

\[
\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \Lambda_{l_1} \subset \Lambda_{l_1+1} \subset \cdots \mathbb{N}_M
\]

and such that the interpolation problem for the points \( \mathcal{Y}_{l_1} \) is well-posed in \( V_{l_1}^\Gamma := \text{span}\{y^\alpha : \alpha \in \Lambda_{l_1}\} \). Naturally, the \( V_{l_1}^\Gamma \) are also nested and we define the interpolation operators

\[
\mathcal{I}_{l_1}^\Gamma : C^0_\chi(\Gamma; H^1_0) \to V_{l_1}^\Gamma
\]

and the difference operators

\[
\Delta^\Gamma_{l_1} := \mathcal{I}_{l_1}^\Gamma - \mathcal{I}_{l_1-1}^\Gamma,
\]

where \( C^0_\chi \) has been defined in (3.6).
Remark 4.3.1. Note that for the index sets $\Lambda_1$, and hence the spaces $V_{1,l}^\Gamma$, to be nested, the sets of collocation points $\mathcal{Y}_{1,l}$ do not necessarily have to share this property. This can easily be seen by considering a one-dimensional Lagrange interpolation operator on $\Gamma_1$, based on the sets $\mathcal{Y}_{1,l}$ of the $l_1 + 1$ abscissae of the Gauss-Legendre quadrature rule of order $l_1$. Evidently, the collocation points are not nested, but the associated spaces $V_{1,l}^\Gamma = \mathcal{P}_{l_1}(\Gamma_1)$ clearly are, where $\mathcal{P}_{l_1}$ denotes the space of polynomials up to order $l_1$.

4.3.2 Sparse tensor stochastic collocation formulation

We first introduce the parametric variational formulation of (4.1): Find $u_M : \Gamma \rightarrow H_0^1(D) : \quad B_M(y; u_M, v) = F(v) \quad \forall v \in H_0^1(D), \quad (4.36)$

where the parametric bilinear form $B_M(y, \cdot, \cdot)$ is defined as

$$B_M(y; u, v) := \int_D a_M(y, x) \nabla u(y, x) \nabla v(x) \, dx, \quad (4.37)$$

and the linear form is given by

$$F(v) = \int_D f(x)v(x) \, dx.$$ 

Note that due to (3.13), the variational problem (4.36) has a unique solution for every fixed $y \in \Gamma$. This also holds true in the presence of a lognormal coefficient, since

$$a^*(y) = e^{\min \mathbb{E}(x) - \sum_{m=1}^{M} \psi_m(x)|y_m|}, \quad a^*(y) = e^{\max \mathbb{E}(x) + \sum_{m=1}^{M} \psi_m(x)|y_m|}. \quad (4.38)$$

Then, we have, for $y \in \Gamma$,

$$B_M(y; u, v) \leq a^*(y)\|u(y, \cdot)\|_{H_0^1(D)} \|v(y, \cdot)\|_{H_0^1(D)},$$
$$B_M(y; u, u) \geq a^*(y)\|u(y, \cdot)\|^2_{H_0^1(D)}. \quad (4.39)$$

Proposition 4.3.2. Let $u_M : \Gamma \rightarrow H_0^1(D)$ be the solution to (4.36), with $a_M$ in (4.37) given either by a linear (4.2) or lognormal (4.3) truncated KL expansion. Then, $u_M \in L^2_p(\Gamma; H_0^1(D))$.

Proof. In the following, we mean by $a^*$ either the lower bound of $a_M$ given in (3.13) or (4.38). By (4.39) for every $y \in \Gamma$ we have

$$\|u_M(y, \cdot)\|^2_{H_0^1(D)} \leq \frac{1}{a^*(y)} B_M(y; u_M, u_M) = \frac{1}{a^*(y)} F(u_M).$$
4.3 Stochastic collocation formulation

\[ \frac{C_P}{a_*(y)} \| f \|_{L^2(D)} \| u_M(y, \cdot) \|_{H^1_0(D)}, \]

and hence

\[ \| u_M(y, \cdot) \|_{H^1_0(D)} \leq \frac{C_P}{a_*(y)} \| f \|_{L^2(D)}. \]

Therefore,

\[ \| u_M \|_{L^2_\rho(\Gamma; H^1_0(D))}^2 = \int_{\Gamma} \| u_M(y, \cdot) \|_{H^1_0(D)}^2 \rho(y) \, dy \leq C_P^2 \| f \|_{L^2(D)}^2 \| \frac{1}{a_*(y)} \|_{L^2_\rho(\Gamma; H^1_0(D))}^2. \]

With the finite dimensional subspaces \( V_{l_2} \) at hand, we solve problem (4.36) for each point \( y \in \mathcal{Y}_{l_1} \) by discretization in the spatial variable \( x \), giving rise to the semidiscrete parametric formulation: Find, for a \( y \in \mathcal{Y}_{l_1} \),

\[ \tilde{u}_M(y, \cdot) \in V_{l_2}^D : \quad B_M(y; \tilde{u}_M(y, \cdot), v) = F(v) \quad \forall v \in V_{l_2}^D. \quad (4.40) \]

By a standard argument, the pointwise quasi-optimality of the Galerkin solution \( u_M \) to (4.40) follows from (4.39):

\[ \| u_M(y, \cdot) - \tilde{u}_M(y, \cdot) \|_{H^1_0(D)}^2 \leq \frac{a_*(y)}{a_*(y)} \inf_{\tilde{v} \in V_{l_2}^D} \| u_M(y, \cdot) - \tilde{v} \|_{H^1_0(D)}^2. \quad (4.41) \]

The random solution \( u_M \in L^2_\rho(\Gamma; V_{l_2}^D) \) is then recovered by interpolating over the collocated solutions by means of (4.34). We can describe the (full) tensor collocation approximation to the solution \( u \) of (4.36) as

\[ u_L := \mathcal{T}_L^\Gamma(P_L^D u) = \sum_{0 \leq l_1, l_2 \leq L} \Delta_{l_2}^\Gamma((P_{l_2}^D - P_{l_2}^{D-1})u), \quad (4.42) \]

where \( P_{l_2}^D \) denotes the \( H^1_0(D) \)-projection (4.32) and \( \Delta_{l_2}^\Gamma \) the incremental interpolation operators (4.35). Consequently, as in (4.23), this approach uses a total number of

\[ N_L = N_L^\Gamma \times N_L^D \]

degrees of freedom where \( N_L^D \) denotes the number of spatial basis functions and \( N_L^\Gamma \) the number of collocation points in the random parameter space, each at level \( L \). Since each collocation point entails one solution of a deterministic problem (4.36), numerous recent works have been focused on reducing the number of collocation points by e.g. a Smolyak construction [43, 42, 67, 23] or ANOVA based construction [56, 9]. In addition to these
techniques, which will be discussed in more detail in Chapter 6, we will also consider a sparse tensor approximation instead of (4.42):

$$\hat{u}_M := \sum_{0 \leq l_1 + l_2 \leq L} \Delta^l_1 ((P^D_{l_2} - P^D_{l_2-1})u).$$  \hfill (4.43)

The total number of degrees of freedom will then, as in Lemma 4.2.1, reduce to essentially log-linear complexity.

**Lemma 4.3.3.** Assume that the number of collocation points between each level \(l_1\) and the dimension of the detail space \(W^D_{l_2}\) do not grow faster than exponential with respect to the levels \(l_1, l_2\), i.e. there exist two geometric sequences \(b^\Gamma_{l_1}\) and \(b^D_{l_2}\), \(l_1, l_2 = 1, 2, 3, \ldots\), with \(b^\Gamma > 1\) and \(b^D > 1\), such that

$$(|Y_{l_1}| - |Y_{l_1-1}|) \lesssim b^\Gamma_{l_1} \quad \text{and} \quad \dim(W^D_{l_2}) \lesssim b^D_{l_2}.$$

Then the sparse tensor approximation of the solution \(u\) to (4.36) by means of (4.43) uses

$$N_L \lesssim L^\theta \max\{b^\Gamma, b^D\}^{L+1}$$

degrees of freedom, where \(\theta = 1\) if \(b^\Gamma = b^D\) and zero otherwise.

**Proof.** The proof is analogous to the one given for Lemma 4.2.1. We observe that

$$N_L \leq c \sum_{l_1=0}^{L} b^\Gamma_{l_1} b^D_{l_2} \leq c \sum_{l_1=0}^{L} b^\Gamma_{l_1} b^D_{l_2-l_1+1}.$$  \hfill (4.45)

In the case where \(b^\Gamma = b^D\), we therefore have

$$N_L \leq c \sum_{l_1=0}^{L} b^{L+1} = c(L + 1)b^{L+1},$$

which leads to (4.44) with \(\theta = 1\). Now we assume w.l.o.g. \(b^D > b^\Gamma\) and it follows from (4.45)

$$N_L \leq c \sum_{l_1=0}^{L} b^\Gamma_{l_1} b^{L-l_1+1}$$

$$= c b^{L+1} \sum_{l_1=0}^{L} b^\Gamma_{l_1} \left(\frac{b^\Gamma}{b^D}\right)^{l_2}$$

$$= c b^{L+1} (b^D/b^\Gamma)^{L+1} - 1 \quad \frac{(b^D/b^\Gamma)^{L+1} - 1}{(b^D/b^\Gamma) - 1}.$$
Since the denominator is positive it can be absorbed in the constant and we find

\[ N_L \leq c b_D^{L+1}. \]

This completes the proof. \(\square\)
4 Sparse tensor discretizations
5 Wavelet discretization in $D$

As we have seen in the previous chapter, a hierarchic sequence of finite-dimensional approximation spaces $V^D_l \subset H^1_0(D)$ is a key component of the sparse tensor stochastic Galerkin and collocation algorithms. In this chapter, we will propose hierarchic finite element wavelets as a basis for the spatial discretization. The construction presented here, is based on [41], which, in principle, allows us to construct wavelets on a regular simplicial triangulation of $D \subset \mathbb{R}^d$. Here, we will treat the case of $d = 1$ and $d = 2$ explicitly. For the case $d = 3$ see [41] and for a more general and detailed introduction into wavelets we refer to [11, 15].

5.1 Wavelet construction

The starting point of the wavelet discretization presented here is based on a nested sequence $\mathcal{T}^l \subset \mathcal{T}^0$ of regular simplicial triangulations of $D$, obtained by successive refinement of an initial mesh $\mathcal{T}_0$. By $\mathcal{I}(\mathcal{T}_l)$ we denote the index set of vertices of the mesh $\mathcal{T}_l$, denoted by $\mathcal{V}(\mathcal{T}_l)$, and by $\hat{\mathcal{I}}(\mathcal{T}_{l+1})$ the index set of vertices of the mesh $\mathcal{T}_{l+1}$, which do not belong to $\mathcal{T}_l$. Hence, $\mathcal{I}(\mathcal{T}_{l+1}) = \mathcal{I}(\mathcal{T}_l) \cup \hat{\mathcal{I}}(\mathcal{T}_{l+1})$, where the union is disjoint, and it therefore holds

$$|\mathcal{I}(\mathcal{T}_l)| = |\mathcal{I}(\mathcal{T}_0)|2^{dl_2}, \quad |\hat{\mathcal{I}}(\mathcal{T}_{l+1})| = |\mathcal{I}(\mathcal{T}_0)|2^{dl_2}(2^d - 1). \quad (5.1)$$

We then define

$$V^D_{l_2} := S^p(D, \mathcal{T}_{l_2}) = \left\{ u \in H^1_0(D) : u|_T \in P_p(T) \text{ for } T \in \mathcal{T}_{l_2} \right\}, \quad (5.2)$$

i.e. the space of continuous piecewise polynomials of degree $p$ w.r.t. the triangulation $\mathcal{T}_{l_2}$. Clearly, the spaces $V^D_{l_2}$ are hierarchic in the sense of (4.17). Denote by $\{\phi^k_{l_2}\}_{k}$, $k \in \nabla_{l_2}$ a basis of $V^D_{l_2} = S^p(D, \mathcal{T}_{l_2})$, with $\nabla_{l_2}$ being a suitable index set.

Given this basis of $V^D_{l_2}$, we then construct a basis $\{\psi^i_{l_2}\}_{i}$, $i \in \nabla_{l_2}$ of the detail spaces $W^D_{l_2}$ (4.19). First, the so-called scaling functions on the level $l_2 = 0$ are defined as the basis functions on the coarsest mesh $\mathcal{T}_0$, i.e. $\psi^0_{l_2} = \phi^k_0$ and $\nabla_0 = \nabla_0$. The construction of wavelets on a higher level $l_2 > 0$ is based on the basis functions on the meshes $\mathcal{T}_{l_2}$ and
\[ T_{l_2-1}. \] First, we construct a family of functions \( \theta^i_{l_2}(x) \in S^p(D, T_{l_2}), \) \( i \in \mathcal{N}_{l_2} \) satisfying \( (\theta^i_{l_2}, \phi^{k_1}_{l_2-1})_{L^2(D)} \approx \delta_{ik}. \) Such functions for \( p = 1 \) are given in \( d = 1 \) by
\[
\theta^i_{l_2}(v) = \begin{cases} 
3 & v = v_i \in \mathcal{V}(T_{l_2-1}) \\
-\frac{1}{2} & v \in \mathcal{V}(T_{l_2}) \text{ is neighbor of } v_i \\
0 & \text{any other } v \in \mathcal{V}(T_{l_2}) 
\end{cases} \quad (5.3)
\]
and in \( d = 2 \) by
\[
\theta^i_{l_2}(v) = \begin{cases} 
14 & v = v_i \in \mathcal{V}(T_{l_2-1}) \\
-1 & v \in \mathcal{V}(T_{l_2}) \text{ is neighbor of } v_i \\
0 & \text{any other } v \in \mathcal{V}(T_{l_2}). 
\end{cases} \quad (5.4)
\]

The ensemble of functions \( \{ \theta^i_{l_2} : i \in \mathcal{I}(T_{l_2-1}) \} \cup \{ \phi^i_{l_2} : i \in \mathcal{I}(T_{l_2}) \} \) forms a \( L^2 \)-Riesz basis of \( V^D_{l_2} \) satisfying \( (\theta^i_{l_2}, \phi^{k}_{l_2-1})_{L^2(D)} = 0 \) if \( i \neq k. \) The wavelets on level \( l_2 \) are then obtained by a Gram-Schmidt orthogonalization process:
\[
\psi^i_{l_2}(x) = \phi^i_{l_2}(x) - \sum_{k \in \mathcal{I}(T_{l_2-1})} \frac{(\phi^i_{l_2}, \phi^{k}_{l_2-1})_{L^2(D)}}{(\theta^i_{l_2}, \phi^{k}_{l_2-1})_{L^2(D)}} \theta^i_{k_{l_2}}(x), \quad i \in \mathcal{I}(T_{l_2}). \quad (5.5)
\]

By construction, \( (\psi^i_{l_2}, \phi^{k}_{l_2-1})_{L^2(D)} = 0, \) and the functions \( \psi^i_{l_2}, i \in \mathcal{N}_{l_2} \) form a uniform \( L^2 \)-Riesz basis for \( W^D_{l_2}, \) see [41, Prop. 3.2.10]. Scaling the wavelets with a factor \( 2^{-l_2}, \) i.e. defining \( \hat{\psi}^i_{l_2}(x) = 2^{-l_2} \psi^i_{l_2}(x), \) then yields a Riesz basis for \( H^1(D). \) In case of homogeneous Dirichlet boundary conditions, the above construction can be modified as follows: For \( v_i \in \mathcal{V}(T_{l_2}), \) the corresponding \( \phi^i_{l_2-1}, \phi^i_{l_2} \) and \( \theta^i_{l_2} \) are excluded from the ensembles and the resulting wavelets \( \psi^i_{l_2}, \) then form a uniform Riesz bases for \( H^1_0(D). \) We refer to Section 5.3, where we will construct the wavelets explicitly in the case \( p = 1, d = 1. \)

### 5.2 Basic properties of wavelet Galerkin approximation

In this section, basic results on the wavelet discretization proposed in the previous section are presented. In particular, we will give an approximation result and a bound on the condition number of stiffness matrices obtained by wavelet discretization.

Recall the \( H^1_0(D) \) Galerkin projection \( P^D_{l_2} : H^1(D) \rightarrow V^D_{l_2} \) from (4.32), where, in the present setting, \( V^D_{l_2} = S^p(D, T_{l_2}). \) We have the following approximation result, which is proved e.g. in [53, Thm. 3.17] for \( d = 1 \) and [52, Prop. 2.32] for \( d = 2, 3. \)
5.2 Basic properties of wavelet Galerkin approximation

Proposition 5.2.1. Assume \( u \in H^{1+t}(D) \) with \( t \in [0,p] \) and \( l_2 \geq 0 \).
\[
\| u - P_{l_2}^D u \|_{H^1_t(D)} \leq C 2^{-l_2} p^{-t} \| u \|_{H^{1+t}(D)}, \quad t \in [0,p],
\] (5.6)
where the constant \( C > 0 \) depends only on the smoothness \( t \) and the quasi-uniformity constant of the mesh \( T_{l_2} \).

Next, the approximation result (5.6) is related to the number of degrees of freedom. Since \( N_{l_2}^D \sim p^{l_2 d} \) we easily derive

Corollary 5.2.2. Assume \( u \in H^{1+t}(D) \) with \( t \in [0,p] \) and \( l_2 \geq 0 \). Then, as \( l_2 \to \infty \),
\[
\| u - P_{l_2}^D u \|_{H^1_t(D)} \lesssim (N_{l_2}^D)^{-t/d} \| u \|_{H^{1+t}(D)}, \quad t \in [0,p].
\] (5.7)

Denote by \( \Psi_{l_2} = \{ \psi_{i,k} \}_{k,i} \) \( k = 0, \ldots, l_2 \) and \( i \in \nabla_{l_2} \) the ensemble of wavelets up to level \( l_2 \), forming a \( H^1_0(D) \) Riesz basis of \( V_{l_2}^D = S^p(D,T_{l_2}) \). Let us further consider the (purely spatial) variational problem of finding a solution \( u \) to \( V_{l_2}^D \)
\[
b(u,v) = l(v) \quad \text{on} \quad D \subset \mathbb{R}^d
\]
with zero Dirichlet boundary conditions, where \( b(\cdot,\cdot) \), in the present context, is a continuous and coercive symmetric bilinear form on \( V_{l_2}^D \times V_{l_2}^D \). Discretizing the problem w.r.t. \( \Psi_{l_2} \) results in a linear system
\[
A u = 1
\] (5.8)
to solve, where the symmetric and positive definite matrix \( A \) is given by \( A = b(\Psi_{l_2},\Psi_{l_2}) \).

Any function \( v \in V_{l_2}^D \) can be written as
\[
v(x) = v^\top \Psi_{l_2} = \sum_j v_j \psi_j(x).
\]

From the uniform Riesz basis property we have the norm equivalence
\[
\| v \|_{H^1_0(D)} \sim \| v^\top \Psi_{l_2} \|_{H^1_t(D)}, \quad \forall v \in V_{l_2}^D
\] (5.9)
with a constant that is independent of the level \( l_2 \). Observing that \( b(u,v) = u^\top A v \) it follows from the continuity and coercivity of \( b(\cdot,\cdot) \) together with (5.9) that
\[
c_1 \leq \frac{v^\top A v}{\| v \|_{\ell_2}^2} \leq c_2 \quad \forall v \in \ell_2
\] (5.10)
with positive constants \( c_1, c_2 \). In other words, the Rayleigh quotient (5.10) of the matrix \( A \) is bounded for any vector \( v \in \ell_2 \), which in turn implies that \( \lambda_{\text{max}}(A) \leq c_2 \) and \( \lambda_{\text{min}}(A) \geq c_1 \). Hence the condition number of the matrix \( A \) with respect to the 2-norm is bounded by
\[
\text{cond}_2(A) \leq c_1 c_2,
\]
independent of the discretization level \( l_2 \). This motivates the use of wavelets, if iterative solvers, such as conjugate gradients, are used to solve the linear system (5.8).
5 Wavelet discretization in $D$

Figure 5.1: The hat functions $\phi^k_{l_2}$, $k \in \mathcal{I}(T_{l_2})$ for $d = 1$ on different levels as indicated, starting from an initial “one-element” mesh $T_0$ with Dirichlet boundary conditions assumed.

5.3 Example

In this section, we will conduct the construction presented in Section 5.1 in the case $d = 1$ and $p = 1$, hence resulting in piecewise linear wavelets. As a basis $\{\phi^k_{l_2}\}_{k \in \nabla(l_2)}$ for the discretization spaces $V^D_{l_2} = S^1(D, T_{l_2})$, we take the standard hat functions, i.e. the piecewise linear polynomials with value 1 at the vertex $k$ and zero at the other nodes, see Figure 5.1. We assume that $T_{l_2}$ consists of equivalent elements, i.e. intervals, with meshwidth $h_{l_2}$ and that the nodes are numbered in ascending order from left to right, i.e. $v_0$ lies on the left boundary of the interval and $v_{N_{l_2}}$ on the right boundary, where $N_{l_2}$ denotes the number of elements of $T_{l_2}$.

Given the functions $\theta^j_{l_2}$ from (5.3), see also Figure 5.2, we easily compute that

$$(\theta^j_{l_2}, \phi^k_{l_2-1})_{L^2(D)} = 2h_{l_2}\delta_{ik} \quad \text{and} \quad (\phi^j_{l_2}, \phi^k_{l_2-1})_{L^2(D)} = \frac{h_{l_2}}{2}$$

if $v_j \in \mathcal{N}(T_{l_2})$ is a neighbor of the vertex $v_k \in \mathcal{N}(T_{l_2-1})$.

From formula (5.5) it then follows that for points $v_i \in \mathcal{N}(T_{l_2})$, which are not adjacent
5.3 Example

Figure 5.2: The hat functions $\phi_{l_2}^k$ on level $l_2$ (top) and $\theta_{l_2+1}^i$ on the finer level $l_2 + 1$ (bottom)

to the boundary ($2 \leq i \leq N_{l_2} - 2$) that

$$\psi_{l_2}^i(v) = \begin{cases} 
\frac{5}{8} & v = v_i \\
-\frac{3}{8} & v = v_{i+1} \in \mathcal{V}(T_{l_2}) \\
\frac{1}{8} & v = v_{i+2} \in \mathcal{V}(T_{l_2}) \\
0 & \text{any other } v \in \mathcal{V}(T_{l_2}) 
\end{cases}$$

For the points $v_i \in \mathcal{N}(T_{l_2})$ adjacent to the boundary, we obtain the boundary wavelets

$$\psi_{l_2}^i(v) = \begin{cases} 
\frac{2}{4} & v = v_1 \text{ or } v = v_{N_{l_2} - 1} \\
-\frac{3}{4} & v = v_2 \text{ or } v = v_{N_{l_2} - 2} \\
\frac{1}{4} & v = v_3 \text{ or } v = v_{N_{l_2} - 3} \\
0 & \text{any other } v \in \mathcal{V}(T_{l_2}) 
\end{cases}$$

Figure 5.3 shows the wavelets and boundary wavelets forming a $L^2(D)$ Riesz basis for $W_{l_2}^D$.

The same construction in the case $p = 1$ and $d = 2$ yields wavelets as shown in Figure 5.4.
5 Wavelet discretization in $D$

Figure 5.3: Top: Hat function basis $\{\phi_k^{l_2}\}_k$ for $V_{l_2}^D$. Middle: Wavelet basis $\{\psi_i^{l_2+1}\}$ of $W_{l_2}^D$. Bottom: boundary wavelets in the case of Dirichlet boundary conditions.

Figure 5.4: Left: The function $\theta_i^{l_2}$ as defined in (5.4). Right: The corresponding piecewise linear finite element wavelet on a triangular mesh in 2-d.
6 Sparse tensor stochastic collocation

We recall from Section 4.3 that the sparse tensor stochastic collocation method (4.43) is based on a hierarchic sequence of discretization spaces \( \{V^D_l\}_l \subseteq \mathbb{R}^d \) (4.17) and a hierarchic sequence of collocation operators \( I_l^T \) (4.34). After the choice of the deterministic discretization spaces \( V^D_l \) was subject of Chapter 5, we are now concerned with the choice of the stochastic collocation operators

\[ I_l^T : C^0_\chi(\Gamma; H^1_0(D)) \to V^T_l \otimes H^1_0(D), \]

with the space \( C^0_\chi(\Gamma; H^1_0(D)) \) as defined in (3.6).

One possible choice is a full tensor Lagrange interpolation operator: Given sequences \( \{y_{m,0}, \ldots, y_{m,\mu_m}\} \subseteq \Gamma_m, \mu_m(l_1) \in \mathbb{N}, 1 \leq m \leq M \), we denote by

\[ Y^\text{full}_{l_1} := \{y_j = (y_{1,j_1}, \ldots, y_{M,j_M}) : 0 \leq j_m \leq \mu_m(l_1), 1 \leq m \leq M\} \quad (6.1) \]

the so-called full tensor grid of collocation points. Then the corresponding full tensor Lagrange interpolation operator is defined as

\[ I^\text{full}_{l_1} v = \sum_{y_j \in Y^\text{full}_{l_1}} v(y_j, x) \ell_j(y), \quad (6.2) \]

where \( \ell_j(y) \) denotes the tensorized Lagrange interpolation polynomial at the point \( y_j \). The application of this interpolation formula, however, requires the evaluation of \( v(\cdot, x) \) at \( N^T = \prod_{m=1}^M (\mu_m + 1) \) points, which, as we have seen in Chapter 4, corresponds to \( N^T \) deterministic PDEs to be solved. In cases where \( M \) is large, this can be prohibitive, due to the so-called ‘curse of dimension’, meaning that \( N^T \) grows exponentially w.r.t. \( M \).

6.1 Smolyak’s construction of collocation points

To keep the number of collocation points moderate, following [43, 42, 67], we will use a Smolyak-type construction. Contrary to those works, however, we will focus on Gaussian instead of Clenshaw-Curtis quadrature rules as the underlying univariate numerical
integration schemes, since their approximation properties allow us to avoid the introduction of a Lebesgue constant in each subspace \( \Gamma_m \), resulting in an overall constant which would depend exponentially on \( M \), see e.g. [4, 43, 42]. Moreover, by renouncing nestedness, we gain the flexibility to increase the number of collocation points in each direction linearly w.r.t. the level of refinement instead of doubling it in each step.

For any \( 1 \leq m \leq M \), any real number \( \gamma \) and a level \( k_m \), let \( \{ y_{m,0}, \ldots, y_{m,\mu_m} \} \subset \Gamma_m \) be the abscissae of the Gauss quadrature rule of order \( \mu_m(k_m) \) with respect to the probability density \( \rho_m \), where

\[
\mu_m(k_m) = \lceil \gamma k_m \rceil.
\]

The parameter \( 0 < \gamma \in \mathbb{R} \) serves as a steering parameter for the growth of the number of collocation points. It allows us to precisely tailor the stochastic refinements to the resolution of the wavelets in space. Its precise value will be determined later in (6.29), based on the expected convergence rates of the spatial discretization.

The one-dimensional interpolation operators \( I_{k_m}^{(m)}(v) \) are then given by

\[
I_{k_m}^{(m)}(v) = \sum_{j_m=0}^{\mu_m(k_m)} v(y_{m,j_m}) \ell_{j_m}(y_m),
\]

where \( \ell_{j_m} \) denotes the Lagrange interpolation polynomial of degree \( \mu_m(k_m) \) in the point \( y_{m,j_m} \), defined by

\[
\ell_{m,j_m}(y_m) \in P_{\mu_m}(\Gamma_m), \quad \ell_{m,j_m}(y_{m,i_m}) = \delta_{j_m,i_m}, \quad 0 \leq j_m, i_m \leq \mu_m.
\]

**Remark 6.1.1.** Standard probability density functions, such as Gaussian or uniform ones, lead to well-known abscissae which are tabulated to full accuracy. For non-standard densities we refer to [24], Theorem 3.1 where an algorithm for the derivation of their associated Gauss nodes and weights is described.

Next, we define the univariate differences

\[
\Delta_{k_m}^{(m)} := I_{k_m}^{(m)} - I_{k_m-1}^{(m)}, \quad \Delta_0^{(m)} = I_0^{(m)}.
\]

Using the multiindex notation \( k = (k_1, \ldots, k_M) \), the (isotropic) \( M \)-dimensional Smolyak interpolation operator is then defined by

\[
I_{l_1}^\Gamma := \sum_{0 \leq |k| \leq l_1} (\Delta_{k_1}^{(1)} \otimes \cdots \otimes \Delta_{k_M}^{(M)})
\]

or, equivalently, by

\[
I_{l_1}^\Gamma = \sum_{0 \leq |k| \leq l_1} (-1)^{|k|} |k|! \binom{M-1}{l_1-|k|} (I_{k_1}^{(1)} \otimes \cdots \otimes I_{k_M}^{(M)}),
\]

50
6.2 Error analysis of Smolyak’s collocation algorithm

For the error analysis of the Smolyak collocation algorithm described above, we assume for now that for each point \(y_j\) the associated elliptic problem (4.36) can be solved exactly in \(H^1_0(D)\). The numerical analysis of the respective full discretization is subject of the next section.

We first provide a result on the approximation properties of the univariate interpolation operators (6.4):

**Proposition 6.2.1.** If \(v\) solves (4.5) and \(I^{(m)}_{k_m}\) denotes the operator defined in (6.4), then the interpolation error admits the following coordinate-wise bound:

\[
\|v - I^{(m)}_{k_m} (v)\|_{L^2_{\mathcal{P}_m} (\Gamma_m; H^1_0(D))} \le C(r_m) e^{-r_m k_m} \|v\|_{C^0_{\mathcal{P}_m} (\Sigma(\Gamma_m, \tau_m); L^2_{\mathcal{R}_m} (\Gamma_m; H^1_0(D)))}
\]

6.2 Error analysis of Smolyak’s collocation algorithm

Figure 6.1: Number of collocation points in a full tensor approach (6.1) compared to the number of collocation points in the Smolyak approach (6.7) for dimensions \(M = 10, 20, 40\)

see [64].

Figure 6.1 illustrates the number of collocation points of the Smolyak approach (6.7) compared to the number of collocation points in a full tensor approach (6.1).
where \( \|v\|_{C^0(\Sigma(\Gamma_m, \tau_m); V)} = \sup_{z \in \Sigma(\Gamma_m, \tau_m)} \chi_m(\text{Re } z) \|v\|_V \) denotes the complex extension of the norm and

- if \( \Gamma_m \) is bounded
  \[
  \begin{align*}
  r_m &= \ln \left( \frac{2\delta_m}{|\Gamma_m|} + \sqrt{1 + \left( \frac{2\delta_m}{|\Gamma_m|} \right)^2} \right) \\
  C(r_m) &= C_1 \frac{1}{|\Gamma_m|}
  \end{align*}
  \]

- if \( \Gamma_m \) is unbounded
  \[
  \begin{align*}
  r_m &= \ln(\sqrt{2\delta_m \tau_m}) \\
  C(r_m) &= C_2(\alpha_m, \delta_m) \frac{1}{|\Gamma_m|}
  \end{align*}
  \]

where \( C_1 \) is independent of \( m, k_m, r_m \) and \( C_2 = e^{\alpha_m/(4\delta_m^2)(1 + \text{Erf}\left( \frac{\alpha_m}{2\delta_m} \right))} \), where \( \tau_m \) defined as in Lemma 3.3.3 and \( \delta_m, \alpha_m \) given by Assumption 3.2.6. \( \text{Erf} \) denotes the Gauss error function, see e.g. [1] Chapter 7.

We note here that the result of the bounded case has already been proposed in [4], but in the unbounded case, Proposition 6.2.1 provides a much sharper bound. In particular we see in the above proposition that in both cases, the constants are decaying to zero since the size of the analyticity regions \( \tau_m \) is growing to infinity as \( m \to \infty \), see also Lemma 3.3.3. The proof of Proposition 6.2.1 will be given at the end of this chapter.

Following the lines of [64, 44] we will now estimate the interpolation error of the Smolyak operator (6.7). First we note that

\[
\mathcal{T}_{l_1}^\Gamma = \sum_{0 \leq |k^*| \leq l_1} \Delta_{k_1}^{(1)} \otimes \cdots \otimes \Delta_{k_{M-1}}^{(M-1)} \otimes \Delta_{k_M}^{(M)}
= \sum_{0 \leq |k^*| \leq l_1} \Delta_{k_1}^{(1)} \otimes \cdots \otimes \Delta_{k_{M-1}}^{(M-1)} \otimes I_{l_1-[k^*]},
\]

(6.10)

where \( k^* = (k_1, \ldots, k_{M-1}) \). In the following denote by \( \text{Id}^{(m)} \) the identity operator on \( \Gamma_m \) and \( \text{Id}^M = \prod_{m=1}^M \text{Id}^{(m)} \). The interpolation error can be written as

\[
\text{Id}^M - \mathcal{T}_{l_1}^\Gamma = (\text{Id}^{M-1} - \mathcal{T}_{l_1}^{\Gamma^*_1}) \otimes \text{Id}^{(M)} + \sum_{0 \leq |k^*| \leq l_1} \Delta_{k_1}^{(1)} \otimes \cdots \otimes \Delta_{k_{M-1}}^{(M-1)} \otimes (\text{Id}^{(M)} - \mathcal{T}_{l_1-[k^*]}^{(M)}),
\]

(6.11)

where \( \Gamma^*_1 = \Gamma_1 \times \ldots \times \Gamma_{M-1} \). In the following, unless otherwise stated, denote by \( \| \cdot \| \) the \( L^2(\Gamma; H^1(D)) \)-norm or its associated operator norm. From

\[
\|\Delta_{k_m}^{(m)}\| \leq \|\text{Id}^{(m)} - I_{k_m}^{(m)}\| + \|\text{Id}^{(m)} - \mathcal{T}_{k_m-1}^{(m)}\| \leq 2C_m e^{-\tau_m \gamma k_m},
\]

(6.12)

where \( C_m = C(r_m) \) as in Proposition 6.2.1, it follows together with (6.11)

\[
\|\text{Id}^M - \mathcal{T}_{l_1}^\Gamma\| \leq \|\text{Id}^{M-1} - \mathcal{T}_{l_1}^{\Gamma^*_1}\| + \sum_{0 \leq |k^*| \leq l_1} \|\Delta_{k_1}^{(1)}\| \cdots \|\Delta_{k_{M-1}}^{(M-1)}\| \|\text{Id}^{(M)} - \mathcal{T}_{l_1-[k^*]}^{(M)}\|
\]

52
\[ \begin{align*}
\leq & \|\text{Id}^{M-1} - T\Gamma_1^{l_1}\| \\
+ & \sum_{0 \leq |k^*| \leq l_1} \left( \prod_{m=1}^{M-1} 2C_me^{-r_mk_m} \right) CMe^{-r_M\gamma_{kM}} \\
= & \|\text{Id}^{M-1} - T\Gamma_1^{l_1}\| + \sum_{0 \leq |k^*| \leq l_1} \prod_{m=1}^{M} 2C_me^{-r_mk_m}.
\end{align*} \]

Setting \( r_{\min} = \min\{r_1, \ldots, r_M\} \) we arrive at the following recursive estimate

\[ \|\text{Id}^{M} - T\Gamma_1^{l_1}\| \leq \|\text{Id}^{M-1} - T\Gamma_1^{l_1}\| + \left( \prod_{m=1}^{M} 2C_me^{-r_{\min}\gamma_{l_1}} \right) \left( M + l_1 \right). \]

We observe from Proposition 6.2.1 that \( C_m \to 0 \) if \( m \to \infty \), hence the product in (6.14) can be bounded from above by a generic constant \( \bar{C} \) which is independent of \( M \). Using (6.13) repeatedly we arrive at

\[ \|\text{Id}^{M} - T\Gamma_1^{l_1}\| \leq \|\text{Id}^{(1)} - T\Gamma_1^{l_1}\| + \frac{1}{2} \sum_{m=2}^{M} \bar{C}e^{-r_{\min}\gamma_{l_1}} \left( M + l_1 \right). \]

Next, we want to estimate the number of collocation points used.

**Lemma 6.2.2.**

\[ N\Gamma_1^{l_1} = \sum_{|k| \leq l_1} \prod_{m=1}^{M} (1 + \gamma k_m) = \left( \gamma l_1 + 2M \right)^2. \]

**Proof.** We first note that

\[ N\Gamma_1^{l_1} = \sum_{|k| \leq l_1 + M} \prod_{m=1}^{M} k_m. \]

Now define \( \alpha_j = \sum_{|k| = j} \prod_{m=1}^{M} k_m \) and observe that

\[ \sum_{j=0}^{\infty} \alpha_j x^j = (x + 2x^2 + 3x^3 + \ldots)^M. \]

The generating function of the series \( a_k = k \) for \( k \geq 1 \) is \( \frac{x}{(1-x)^2} \), see [66], hence

\[ \sum_{j=0}^{\infty} \alpha_j x^j = \left( \frac{x}{(1-x)^2} \right)^M. \]
This in turn means that $\alpha_j$ is the $(j - M)$-th coefficient of the power series of $\frac{1}{(1-x)^{2M}}$, which, by [66] Section 2.5, is equal to $\binom{j+M-1}{2M-1}$. Summing the $\alpha_j$'s therefore leads to

$$\sum_{|k| \leq \gamma l_1 + M} \prod_{m=1}^{M} k_m = \sum_{j=0}^{\gamma l_1 + M} \alpha_j = \binom{\gamma l_1 + 2M}{2M}, \quad (6.20)$$

which completes the proof. \square

Inserting the estimate for $N_{l_1}^\Gamma$ into (6.15) we obtain

**Lemma 6.2.3.** Let $u_M$ be the solution to (4.5). Then the collocation error admits the following bounds with respect to the level $l_1$ and the number of collocation points:

$$\|u_M - \mathcal{I}_{l_1}^\Gamma u_M\|_{L^2_2(\Gamma; H^{1}_{0}(D))} \leq C e^{-r_{\min} \gamma l_1} \left( \frac{M + l_1}{M} \right) \|u_M\|_{C^0_h(\Sigma(\Gamma, \tau); H^{1}_{0}(D))}$$

and

$$\|u_M - \mathcal{I}_{l_1}^\Gamma u_M\|_{L^2_2(\Gamma; H^{1}_{0}(D))} \leq C \sqrt{N_{l_1}^\Gamma} \exp\left( -r_{\min}(N_{l_1}^{1/(2M)} - 1) \frac{2M}{1 + \ln 2M} \right) \times \|u_M\|_{C^0_h(\Sigma(\Gamma, \tau); H^{1}_{0}(D))},$$

where $C > 0$ is independent of $M$ and $N_{l_1}^\Gamma$.

**Proof.** From (6.16) and using the inequality between geometric and arithmetic mean, we have

$$N_{l_1}^\Gamma \leq \binom{\gamma l_1 + 2M}{2M} \leq \binom{\gamma l_1 + 2M}{2M} \frac{(\gamma l_1 + 2M)(\gamma l_1 + 2M - 1) \cdots (\gamma l_1 + 1)}{(2M)!} \leq \binom{\gamma l_1 + 2M}{2M} \leq \left( \frac{2M + \gamma l_1 \sum_{m=1}^{2M} 1/m}{2M} \right)^{2M} \leq \left( 1 + \frac{\gamma l_1 (1 + \ln 2M)}{2M} \right)^{2M}. \quad (6.23)$$

Hence, $l_1 \geq ((N_{l_1}^\Gamma)^{1/2M} - 1) \frac{2M}{\gamma (1 + \ln 2M)}$ from which (6.22) follows. \square
6.3 Error Analysis

Remark 6.2.4. As it can be seen in Lemma 6.2.3, the convergence rates with respect to the number of collocation points are exponential but depending on $M$. Since $M$ can possibly be very large in applications, it is of interest to examine this case in some more detail. From (6.23) we have for large $M$

$$N_{l_1}^\Gamma \leq e^{\gamma l_1 (1 + \ln 2M)}$$  \hspace{1cm} (6.24)

which, inserted into (6.21) and noting that

$$\left(\frac{l_1 + M}{M}\right) \leq \left(\frac{(l_1 + M)e}{M}\right)^M \leq e^M \left(1 + \frac{l_1}{M}\right)^M \leq e^{l_1 + M}$$

gives

$$\|u_M - I_{l_1,1}^\Gamma u_M\|_{L_2^\Gamma (\Gamma; H^1_0(D))} \leq C(N_{l_1}^\Gamma)^{-(r_{\min} - 1)/(1 + \ln 2M)} \|u_M\|_{C^0(\Sigma(\Gamma, \tau); H^1_0(D))}.$$  

Hence, for $M \gg 1$, the exponential convergence rate in Lemma 6.2.3 behaves algebraically in the preasymptotic range, i.e. for $l_1 \ll M$.

Remark 6.2.5. So far, we only considered an isotropic Smolyak algorithm, i.e. with an equal number of quadrature points in each direction. However, one could also consider an anisotropic Smolyak algorithm, i.e. with

$$\mu_m(k_m) = \lceil \gamma \xi_m k_m \rceil$$  \hspace{1cm} (6.25)

instead of (6.3), where $\{\xi_m\}_{m=1}^M \in \mathbb{R}^M$ is a sequence of weights. Hence, the isotropic formula is the special case $\xi_m = 1, \ m = 1, \ldots, M$. The (anisotropic) Smolyak operator $I_{l_1,1}^\Gamma$ (6.7) is then defined analogously by replacing (6.3) with (6.25), featuring a hierarchy w.r.t. the level $l_1$. Choosing the weights $\xi_m = 1/r_m, \ r_m$ given by Proposition 6.2.1, we obtain instead of (6.15)

$$\|\text{Id}^M - I_{l_1,1}^\Gamma \| \leq \tilde{C} e^{-\gamma l_1 \left(\frac{M + l_1}{M}\right)}$$

The techniques used in Lemma 6.2.2, however, to compute the number of collocation points, do not work any more in the anisotropic case, since it is no longer a problem of counting integers.

6.3 Error Analysis

The main focus of this section aims at proving convergence rates for the sparse tensor stochastic collocation method. We note that these results hold for both, bounded and unbounded domains $\Gamma_m$. We assume in the following that the spatial discretization is given by piecewise polynomials of degree $p$ on a quasi-uniform mesh of width $h$, as defined in (5.2).
Proposition 6.3.1. Let $u_M$ be the solution to (4.5) and denote by $\hat{u}_M$ its sparse tensor approximation given by (4.43). We have the following error estimate

$$\|u_M - \hat{u}_M\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq C N_L^{-\beta} \|u_M\|_{C_0(\Sigma(\Gamma, \tau): H^{1+c} \cap H^1_0(D))}$$

(6.26)

in terms of the total number of degrees of freedom $N_L$ with rate

$$\beta(M) = \min\{\frac{r_{\min} - 1}{1 + \ln 2 M}, \frac{t \ln 2}{d}\}$$

(6.27)

and $t \in [0, p]$.

In other words, by the sparse composition of hierarchic interpolation operators and a multilevel basis in space we retrieve the smaller of the two convergence rates provided by the spatial discretization (5.7) and stochastic interpolation, see Remark 6.2.4.

Remark 6.3.2 (Full tensor collocation). The main motivation for the sparse tensor composition (4.43) of stochastic interpolation operators and spatial Galerkin projectors lies in the reduction of degrees of freedom to $O(L^\theta \max\{N_L^{\Gamma}, N_L^D\})$ as shown in Lemma 4.2.1, opposed to $O(N_L^{\Gamma} \times N_L^D)$ in the full tensor approach (4.42). Using the approximation results (5.6) and (6.21), we can easily derive the full tensor error estimate

$$\|u_M - I_{L, \Gamma}^\Gamma(P_D^L u_M)\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq C(N_L)^{-\beta} \|u_M\|_{C_0(\Sigma(\Gamma, \tau): H^{1+c} \cap H^1_0(D))}$$

(6.28)

with rate $\bar{\beta} = (d/t + (1 + \ln 2 M)/(r_{\min} - 1))^{-1}$.

Proof. First, we will estimate the approximation error in terms of the global refinement level $L$. To simplify the notation, we will drop the subscript in the norms, hence $\| \cdot \| = \| \cdot \|_{L^2_\rho(\Gamma; H^1_0)}$ unless otherwise indicated. Using (4.43) we obtain

$$\frac{\|u_M - \hat{u}_M\|}{\|u_M\|} = \frac{\left(\sum_{l_1=0}^L (\sum_{l_2=L-l_1+1}^{\infty} \Delta_{l_1} (P_{l_2}^{D} - P_{l_2-1}^{D}) u_M)\right)}{\left(\sum_{l_1=0}^L (\sum_{l_2=L+1}^{\infty} \Delta_{l_1} (P_{l_2}^{D} - P_{l_2-1}^{D}) u_M)\right)}$$

(6.29)

$$+ \frac{\left(\sum_{l_1=L+1}^{\infty} \sum_{l_2=0}^{\infty} \Delta_{l_1} (P_{l_2}^{D} - P_{l_2-1}^{D}) u_M\right)}{\left(\sum_{l_1=0}^L (\sum_{l_2=L-l_1+1}^{\infty} \Delta_{l_1} (P_{l_2}^{D} - P_{l_2-1}^{D}) u_M)\right)}$$

(6.30)

$$= \frac{\left(\sum_{l_1=0}^L \Delta_{l_1} (Id^D - P_{L-l_1}^D) u_M\right)}{\left(\sum_{l_1=0}^L (\sum_{l_2=L-l_1+1}^{\infty} \Delta_{l_1} (P_{l_2}^{D} - P_{l_2-1}^{D}) u_M)\right)}$$

(6.31)

$$+ \frac{\left(\sum_{l_1=L+1}^{\infty} \sum_{l_2=0}^{\infty} \Delta_{l_1} (P_{l_2}^{D} - P_{l_2-1}^{D}) u_M\right)}{\left(\sum_{l_1=0}^L (\sum_{l_2=L-l_1+1}^{\infty} \Delta_{l_1} (P_{l_2}^{D} - P_{l_2-1}^{D}) u_M)\right)}$$

(6.32)

$$\leq \sum_{l_1=0}^L \| (Id^D - I_{l_1}^\Gamma) (Id^D - P_{L-l_1}^D) u_M \|$$

(6.33)
6.4 Proof of Proposition 6.2.1

\[ + \sum_{l_1=0}^{L} \|(1d - T_{l_1}^D)(1d - P_{L-l_1}D)u_M\| + \|(1d - T_{l_1}^D)1d u_M\| \]

\[ \leq \left( \sum_{l_1=0}^{L} C \left( \frac{M + l_1}{M} \right) e^{-\gamma_{\text{min}} l_1} e^{-\ln 2(L-l_1)} \right) \]

\[ + C \left( \frac{M + L}{M} \right) e^{-\gamma_{\text{min}} L} \|u_M\|_{C^0_{\chi}(\Sigma(\Gamma,\tau);H^{1+\mu} \cap H^1(D))}, \]

where we used the estimates (5.6) and (6.21) in the last line. Absorbing now the binomial factor into the exponent by

\[ \left( \frac{M + l_1}{M} \right) \leq \left( \frac{M + l_1}{M} \right)^M \leq e^{M} \left( 1 + \frac{l_1}{M} \right)^M \leq e^{l_1 + M} \]

and choosing

\[ \gamma = \frac{t \ln 2}{r_{\text{min}} - 1} \quad (6.29) \]

we arrive at

\[ \|u_M - \hat{u}_M\|_{L^2(\Gamma;H^1(D))} \leq C e^{-(\ln 2)L} \|u_M\|_{C^0_{\chi}(\Sigma(\Gamma,\tau);H^{1+\mu} \cap H^1(D))} \quad (6.30) \]

Next, we estimate the total number of degrees of freedom. We have from (6.24) that

\[ N_{l_1} \leq \exp\left( \frac{t \ln 2}{r_{\text{min}} - 1} (1 + \ln 2M)l_1 \right), \]

which satisfies the assumptions of Lemma 4.2.1 by setting \( b_D = 2^d \), we obtain

\[ N_L \leq \max\{2^d, 2^{(1+\ln 2M)/(r_{\text{min}} - 1)}\}^L \quad (6.31) \]

The proof is completed by combining (6.31) together with (6.30).

\[ \Box \]

6.4 Proof of Proposition 6.2.1

In the following, to keep the results as general as possible, denote by \( \mathcal{W}(D) \) a Banach space. It has already been proved in [4] (cf. Lemma 4.3) that the \( L^2_{\rho_m}(\Gamma_m) \) interpolation error based on Gaussian abscissae can be bounded from above by a best-approximation error, i.e.

**Lemma 6.4.1.** For every function \( v \in C^0_{\chi_m}(\Gamma_m;\mathcal{W}(D)) \), the interpolation error satisfies

\[ \|v - I_{\mu_m}^{(m)}(v)\|_{L^2_{\rho_m}(\Gamma_m;\mathcal{W}(D))} \leq C \inf_{w \in P_{\mu_m}(\Gamma_m) \otimes \mathcal{W}(D)} \|v - w\|_{C^0_{\chi_m}(\Gamma_m;\mathcal{W}(D))}, \]

where \( C > 0 \) is independent of \( \mu_m \) and \( I_{\mu_m}^{(m)} \) is given by (6.4) based on the zeros of orthogonal polynomials w.r.t. \( \rho_m \).
It therefore remains to bound the best approximation error in the norm $C^0$. To this end, we distinguish the cases where $\Gamma_m$ is bounded or unbounded. For the bounded case we refer to [4], Lemma 4.4.

Lemma 6.4.2. For every function $v \in C^0([-1,1], W(D))$ which admits an analytic extension to the region $\Sigma([-1,1], \tau_m) \subset \mathbb{C}$ as in (3.9) for some $\tau_m > 0$, it holds that

$$
\inf_{w \in P_{\mu_m}([-1,1]) \otimes W(D)} \|v - w\|_{C^0([-1,1], W(D))} \leq \frac{2}{e^{\tau_m} - 1} e^{-\mu_m \tau_m} \max_{z \in \Sigma([-1,1], \tau_m)} \|v(z)\|_{W(D)}
$$

for $r_m = \ln(\tau_m + \sqrt{1 + \tau_m^2})$.

By rescaling the domain $[-1,1]$ to $\Gamma_m$, Proposition 6.2.1 therefore follows in the bounded case by combining Lemma 6.4.1 and 6.4.2 with $\mu_m = \lceil \gamma k_m \rceil$.

In the unbounded case we will, as in [4], first refer to a result from [31] but then proceed in a slightly different way to obtain the result claimed in Proposition 6.2.1. We denote by $H_n$ the Hermite polynomial of order $n$:

$$
H_n(t) = (-1)^n e^{t^2/2} \frac{\partial^n}{\partial t^n} e^{-t^2/2}, \quad t \in \mathbb{R}
$$

satisfying

$$
\int_{\mathbb{R}} H_m(t) H_n(t) e^{-t^2/2} dt = \sqrt{2\pi n!} \delta_{mn}.
$$

Furthermore we define by $h_n(t) = e^{-t^2/4} \frac{1}{\sqrt{\sqrt{2\pi n!}}} H_n(t)$ the $L^2(\mathbb{R})$-orthonormal Hermite functions. The following Lemma from [31] (cf. Theorem 1) then provides a necessary and sufficient condition for Fourier-Hermite series to converge in a strip around the real axis.

Lemma 6.4.3 (Hille 1940). Let $f : \Sigma(\mathbb{R}, \tau_m) \to \mathbb{C}$ be analytic. A necessary and sufficient condition in order that the Fourier-Hermite series

$$
\sum_{k=0}^{\infty} f_k h_k(z_m), \quad f_k = \int_{\mathbb{R}} f(y_m) h_k(y_m) dy_m
$$

exists and converges to the sum $f(z_m)$ in $\Sigma(\mathbb{R}, \tau_m)$, is that for every $\beta_m \in [0, \tau_m)$ there exists a finite constant $C(\beta_m) > 0$, such that

$$
|f(y_m + iw_m)| \leq C(\beta_m) e^{-|y_m| \sqrt{\beta_m^2 - w_m^2}}, \quad y_m \in \mathbb{R}, \; w_m \in [-\beta_m, +\beta_m].
$$
Recall the measure $\chi_m(y_m) = e^{-\alpha_m |y_m|}$ from (3.7) and furthermore define the Gaussian measure $G_m(y_m) = e^{-(\delta_m y_m)^2}$ for $\delta_m$ as in Assumption 3.2.6. Note that Lemma 6.4.1 holds for both of them, see [4] for details. By $v : \Gamma_m \rightarrow \mathbb{R}$ denote a function which is analytic in the strip $\Sigma(\mathbb{R}, \tau_m)$.

In order to apply Lemma 6.4.3, we have to rescale the measure $G$ to the standard Gaussian measure by the change of variables $t_m = \sqrt{2} \delta_m y_m$. We then define $\tilde{v}(t_m) = v(y_m(t))$. Since $\tilde{v}$ is the composition of two analytic functions, it is itself again an analytic function and its domain of analyticity rescales to $\Sigma(\mathbb{R}, \sqrt{2} \tau_m \delta_m)$.

Next, we show that the formal Hermite series of $\tilde{v}$

$$\tilde{v}(t_m) = \sum_{k=0}^{\infty} v_k H_k(t_m),$$

with coefficients $v_k \in \mathcal{W}(D)$ given by

$$v_k = \frac{1}{\sqrt{2\pi k!}} \int_{\mathbb{R}} \tilde{v}(t_m) H_k(t_m) e^{-t_m^2/2} dt_m$$

is well-defined and converges in $\Sigma(\mathbb{R}, \sqrt{2} \tau_m \delta_m)$. To be in the setting of Lemma 6.4.3, we set $f(z_m) = \tilde{v}(z_m) e^{-z_m^2/4}$ and observe that

$$f_k = \int_{\mathbb{R}} f(t_m) h_k(t_m) dt_m = \frac{1}{\sqrt{2\pi k!}} \int_{\mathbb{R}} \tilde{v}(t_m) H_k(t_m) e^{-t_m^2/2} dt_m = \sqrt{2\pi k!} v_k. \quad (6.36)$$

$f$ is clearly analytic in the strip $\Sigma(\mathbb{R}, \sqrt{2} \tau_m \delta_m)$ as it is a product of two analytic functions which are each analytic in $\Sigma(\mathbb{R}, \sqrt{2} \tau_m \delta_m)$. Furthermore, we have that

$$\|f(t_m + iw_m)\|_{\mathcal{W}(D)} = |e^{-\frac{(t_m+iw_m)^2}{4}}\|\tilde{v}(z_m)\|_{\mathcal{W}(D)} \leq e^{-\frac{t_m^2 - w_m^2}{4}} e^{\frac{\alpha}{2 \delta_m t_m}} \|v\|_{C^0(\chi_m(\Sigma(\mathbb{R}, \tau_m); \mathcal{W}(D))}.$$

The function $f$ thus satisfies the hypotheses of Lemma 6.4.3 by setting

$$C(\beta_m) = \max_{t_m \in \mathbb{R}, w_m \in [-\beta_m, \beta_m]} \exp \left\{ -\frac{t_m^2 - w_m^2}{4} + \frac{\alpha}{\sqrt{2} \delta_m} |t_m| + |t_m| \sqrt{\frac{\beta_m^2 - w_m^2}{2}} \right\}.$$ 

Hence, the expansion of $f$ into Hermite functions $h_k$ is converging pointwise in the strip $\Sigma_m(\mathbb{R}, \sqrt{2} \tau_m \delta_m)$ and therefore, by (6.36) and

$$\sum_{k=0}^{\infty} v_k H_k(z_m) = \sum_{k=0}^{\infty} f_k e^{t_m^2/4} h_k(z_m) = e^{t_m^2/4} f(z_m) = \tilde{v}(z_m),$$
also the Hermite series of \( \tilde{v} \).

Next, we will establish a bound for the Hermite coefficients \( v_k \). We define \( \tilde{r}_m := \sqrt{2r_m \delta_m} \) and denote by \( C_{\tilde{r}_m}(\tilde{r}_m) \) the circle around \( t_m \) with radius \( \tilde{r}_m \). Integrating by parts and using Cauchy’s integral formula we obtain

\[
\|v_k\|_{W(D)} = \frac{1}{\sqrt{2\pi k!}} \left| \int_{\mathbb{R}} \tilde{v}(t_m) H_k(t_m) e^{-t_m^2/2} dt_m \right|_{W(D)}
\]

\[
= \frac{1}{\sqrt{2\pi k!}} \left| \int_{\mathbb{R}} \tilde{v}(t_m) \frac{\partial^k}{\partial t_m^k} e^{-t_m^2/2} dt_m \right|_{W(D)}
\]

\[
= \frac{1}{\sqrt{2\pi k!}} \left| \int_{\mathbb{R}} \tilde{v}(t_m) e^{-t_m^2/2} dt_m \right|_{W(D)}
\]

\[
\leq \frac{1}{\sqrt{2\pi \tilde{r}_m^k}} \left| v \right|_{C^0_{\chi_m}(\Sigma(\mathbb{R}, \tilde{r}_m); W(D))} \int_{\mathbb{R}} e^{-t_m^2/2 + \alpha_m \frac{t_m}{\sqrt{2\delta_m}}} e^{-\tilde{r}_m^2/2} dt_m.
\]

Hence, the Fourier Hermite coefficients satisfy the bound

\[
\|v_k\|_{W(D)} \leq C(\alpha_m, \delta_m) \frac{\tilde{r}_m^k}{\sqrt{2\delta_m}} e^{-k \ln \tilde{r}_m} \|v\|_{C^0_{\chi_m}(\Sigma(\mathbb{R}, \tilde{r}_m); W(D))},
\]

where \( C(\alpha_m, \delta_m) = e^{\alpha_m^2/(4\delta_m^2)} \left( 1 + \text{Erf} \left( \frac{\alpha_m}{2\delta_m} \right) \right) \). Therefore, we can estimate the best approximation error in Lemma 6.4.1 by

\[
E_{\mu_m}(v) := \inf_{w \in \mathcal{P}_{\mu_m} \otimes W(D)} \|v - w\|_{C^0_{\chi_m}(\Sigma(\mathbb{R}, \tilde{r}_m); W(D))} \leq \max_{t_m \in \mathbb{R}} \sum_{k=\mu_m+1}^{\infty} v_k h_k(t_m) \|v\|_{W(D)}.
\]

Since \( |h_k(t_m)| < 1 \) for all \( t_m \in \mathbb{R} \), \( k \in \mathbb{N}_0 \) and due to (6.37), the truncated Hermite series can be bounded by

\[
E_{\mu_m}(v) \leq \sum_{k=\mu_m+1}^{\infty} \|v_k\|_{W(D)} \leq C(\alpha_m, \delta_m) \frac{\tilde{r}_m^k}{\sqrt{2\delta_m}} \|v\|_{C^0_{\chi_m}(\Sigma(\mathbb{R}, \tilde{r}_m); W(D))} \sum_{k=\mu_m+1}^{\infty} e^{-k \ln \tilde{r}_m}.
\]

For the series on the r.h.s. we use the summation formula for geometric series to obtain

\[
E_{\mu_m}(v) \leq C(\alpha_m, \delta_m) e^{\tau_m \alpha_m} \frac{e^{-\mu_m \ln \tilde{r}_m} \tilde{r}_m}{r_m - 1}.
\]

From Section 3.3.3 and (3.42) we have that \( \alpha_m \sim \tau_m^{-1} \), and hence the term \( e^{\tau_m \alpha_m} \) remains constant. Setting \( r_m := \ln \tilde{r}_m \), the claim now follows from (6.38) and Lemma 6.4.1.
7 Sparse tensor stochastic Galerkin

We recall from Section 4.2 that we discretize the variational formulation

\[ b_M(\hat{u}_M, v) = l(v), \quad (7.1) \]

\( b_M(\cdot, \cdot) \) and \( l(\cdot) \) given by (4.6) and (4.7), respectively, with respect to the sparse tensor product space \( V^T_L \otimes V^D_L \) where

\[ \{V^T_{l_1}\}_{l_1} \subset L^2_\rho(\Gamma) \quad \text{and} \quad \{V^D_{l_2}\}_{l_2} \subset H^1_0(D) \]

are two hierarchic families of subspaces, see also (4.16) and (4.17). The choice of the spatial discretization spaces \( V^D_{l_2} = S^p(D, \mathcal{T}_{l_2}) \) has been subject of Chapter 5 and we are now concerned with the choice of the stochastic spaces \( V^T_{l_1} \) which is the crucial part in the stochastic Galerkin FEM due to the possibly arbitrary high dimension \( M \) of the parameter domain \( \Gamma \).

7.1 Hierarchic polynomial chaos approximation in \( L^2_\rho(\Gamma; H^1_0(D)) \)

It is therefore imperative to reduce the number of stochastic degrees of freedom to a minimum while still retaining accuracy. To this end, we will, in the spirit of a best-\( N \)-term approximation, approximate the solution by a polynomial chaos expansion in which only the ‘most important’ terms are retained.

7.1.1 Preliminary conventions and notations

In the discussion that follows, we will distinguish the cases where the stochastic parameter domains \( \Gamma_m \) are bounded and unbounded. In the first case, we can w.l.o.g. assume that \( \Gamma_m = [-1, 1], \forall m \). Moreover, we will then restrict our discussion to the case of a uniform probability density \( \rho_m(y_m) = \frac{1}{2} \). In the case of unbounded domains we assume \( \Gamma_m = \mathbb{R} \) and, motivated by Assumption 3.2.6 and after a coordinate transformation, that \( \rho_m(y_m) = \frac{1}{\sqrt{2\pi}} e^{-(y_m)^2/2} \).
By $\mathbb{N}_0^M$, we denote the set of all multiindices of length $M$, i.e. all sequences $\nu = (\nu_1, \nu_2, \ldots, \nu_M)$ of non-negative integers. For $\nu \in \mathbb{N}_0^M$, we denote
$$|\nu|_0 = \#\text{supp}(\nu) \quad \text{and} \quad |\nu| = \sum_{m \geq 1} |\nu_m|,$$
where $\text{supp}(\nu) := \{ j \in \mathbb{N} : \nu_j \neq 0 \}$. For $\nu \in \mathbb{N}_0^M$, we denote the tensorized Legendre polynomial of degree $\nu$ by
$$L_\nu(y) := L_{\nu_1}(y_1)L_{\nu_2}(y_2)\cdots L_{\nu_M}(y_M) \quad y \in \Gamma,$$
where $L_{\nu_m}(y_m)$ denotes the Legendre polynomial of degree $\nu_m$ on $[-1, 1]$ normalized such that $L_{\nu_m}(1) = 1$, i.e. they satisfy the orthogonality relation
$$\int_{-1}^1 L_{\nu_m}(y_m) L_{\mu_m}(y_m) \rho_m(y_m) \, dy_m = \frac{\delta_{\nu_m, \mu_m}}{2^{\nu_m + 1}}.$$
Likewise, we define the tensorized Hermite polynomial of degree $\nu$ by
$$H_\nu(y) := H_{\nu_1}(y_1)H_{\nu_2}(y_2)\cdots H_{\nu_M}(y_M) \quad y \in \Gamma, \nu \in \mathbb{N}_0^M,$$
where $H_{\nu_m}(y_m)$ denotes the Hermite polynomial of degree $\nu_m$ on $\mathbb{R}$ satisfying the orthogonality relation
$$\int_{\mathbb{R}} H_{\nu_m}(y_m) H_{\mu_m}(y_m) \frac{1}{\sqrt{2\pi}} e^{-y_m^2/2} \, dy_m = \nu_m! \delta_{\nu_m, \mu_m}.$$

### 7.1.2 Best-$N$-term polynomial chaos approximation

In the following, to keep the notation simple and to avoid repetitions, we denote by $\mathcal{P}_\nu$ either of the polynomials $L_\nu$ or $H_\nu$, depending on whether we consider the case where the $\Gamma_m$ are bounded or unbounded.

From Section 2.3 we have that $L_\rho^2(\Gamma) = \overline{\text{span}\{\mathcal{P}_\nu : \nu \in \mathbb{N}_0^M\}}$, where the closure is w.r.t $L_\rho^2(\Gamma)$ and the solution $u_M$ to the parametric, deterministic problem (4.36) can be represented in terms of $\rho$-orthogonal polynomials as
$$u_M(y, x) = \sum_{\nu \in \mathbb{N}_0^M} u_\nu(x) \mathcal{P}_\nu(y), \quad (7.8)$$
7.1 Hierarchic polynomial chaos approximation in $L^2_\rho(\Gamma; H^1_0(D))$

where the ‘coefficients’ $u_\nu \in H^1_0(D)$ are given by

$$u_\nu = \left( \prod_{m=1}^{M} 2\nu_m + 1 \right) \int_\Gamma u_M(y, \cdot) L_\nu(y) \rho(y) \, dy$$

in the Legendre case and by

$$u_\nu = \frac{1}{\nu!} \int_\Gamma u_M(y, \cdot) H_\nu(y) \rho(y) \, dy$$

in the Hermite case, where the equalities hold in $H^1_0(D)$.

Given any index set $\Lambda \subset \mathbb{N}_0^M$ of cardinality $|\Lambda| < \infty$, we denote by

$$u_{M,\Lambda}(y, x) = \sum_{\nu \in \Lambda} u_\nu(x) P_\nu(y)$$

the $\Lambda$-truncated orthogonal PC expansion (7.8). In the Galerkin approximation below, we aim at finding an increasing sequence $\Lambda(1) \subset \Lambda(2) \subset \Lambda(3) \subset \cdots \subset \mathbb{N}_0^M$ of index sets $\Lambda(l_1)$, such that the approximations $u_{M,\Lambda(l_1)}$ converge at a certain rate $0 < r \in \mathbb{R}$ in terms of $|\Lambda(l_1)|$.

For exponential eigenvalue decay (2.15) in the KL-expansion (2.8), this is indeed possible with arbitrary high algebraic convergence rates $r$ as it has been shown in [9, 59]. There, an algorithm has been presented to a-priori locate index set $\Lambda$, based only on the input data $a, f$, such that for any prescribed rate $r > 0$, there exists a constant $C(r) > 0$, independent of $l_1$ and $M$ and

$$\|u_M - u_{M,\Lambda(l_1)}\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq C(r)|\Lambda(l_1)|^{-r}.$$  

(7.12)

The case of algebraic decay (2.14) of the Karhunen-Loève eigenvalues however, is more difficult since, as it will turn out, they imply a slow decay of the ‘coefficients’ in (7.11). Here, we use best $N$-term approximations of $u_M$ in (7.8) as a benchmark and prove a result similar to (7.12) in the case of algebraic Karhunen-Loève eigenvalue decay for a certain range of rates $r$, by imposing sufficient conditions on the input data.

To this end, for any finite set $\Lambda \subset \mathbb{N}_0^M$ and for any Banach space $W(D)$, we denote by

$$\Pi(\Lambda; W(D)) := \{v = \sum_{\nu \in \Lambda} v_\nu P_\nu : v_\nu \in W(D), \nu \in \Lambda \} \subset L^2_\rho(\Gamma; H)$$

(7.13)

the linear space of $W(D)$-valued polynomials, which can be expressed as a finite linear combination of $|\Lambda|$ many $\rho(dy)$ orthogonal polynomials $P_\nu$ (if $\Lambda = \emptyset$, we define $\Pi(\Lambda; W(D)) = \{0\}$). They are a countable $\rho(dy)$-orthonormal basis of $L^2_\rho(\Gamma)$ and any
7 Sparse tensor stochastic Galerkin

\( v \in L^2_\rho(\Gamma; W(D)) \) is determined by its coefficient vector \((\nu_\nu : \nu \in \mathbb{N}^M_c) \subset W(D)\) via

\[
v = \sum_{\nu \in \mathbb{N}^M_c} \nu_\nu P_\nu.
\]

The best \(N\)-term semidiscrete approximation error of \(u \in L^2_\rho(\Gamma; W(D))\) is defined as

\[
\sigma_N(u) := \inf_{\Lambda \subset \mathbb{N}^M_c, \vert \Lambda \vert \leq N} \inf_{v \in \Pi(\Lambda; W(D))} \| u - v \|_{L^2_\rho(\Gamma; H)}.
\]

Note that \(\sigma_N(u)\) is uniquely defined even though \(v\) might not be unique.

**Remark 7.1.1.** As is well-known, best \(N\)-term approximations are generally not computationally accessible. We therefore present in Section 8.1 an algorithm to locate for certain sequences \(\{\lambda_m\}_{m=1}^\infty\) of Karhunen-Loève eigenvalues corresponding sequences \(\{\Lambda_l\}_{l=0}^\infty\) of index sets \(\Lambda_l \subset \mathbb{N}^M_0\) of ‘active’ indices in the chaos expansions (7.11) which realize ‘quasi-best \(N\)-term approximations’.

We define \(\mathcal{A}_r(W(D))\) as the set of those \(u \in L^2_\rho(\Gamma; W(D))\), for which

\[
\| u \|_{\mathcal{A}_r(W(D))} := \sup_{N \geq 1} N^r \sigma_N(u)
\]

is finite. This class consists of functions in \(L^2_\rho(\Gamma; W(D))\), which can be best \(N\)-term approximated at a rate \(r\). For a parameter \(\gamma > 0\) we introduce the index sets

\[
\Lambda_\gamma(l_1) := \arg \max_{\Lambda \subset \mathbb{N}^M_c, \vert \Lambda \vert = \lceil 2^\gamma l_1 \rceil} \left( \sum_{\nu \in \Lambda} \| u_\nu \|_{W(D)} \right) \subset \mathbb{N}^M_0, \quad l_1 = 0, 1, 2, \ldots
\]

of the \(\lceil 2^\gamma l_1 \rceil\) coefficients \(u_\nu\) in (7.8), with the largest \(W(D)\)-norm. Thus, \(\Lambda_\gamma(l_1)\) provides a best \(\lceil 2^\gamma l_1 \rceil\)-term approximation of \(u \in L^2_\rho(\Gamma; H)\).

To apply these concepts to (4.4), we choose \(W(D) = H^1_0(D)\). Again, as in the previous chapter, \(\gamma\) serves as a steering parameter for the stochastic resolution and will be determined in Section 7.2, based on the expected convergence rate of the spatial discretization. The corresponding approximation spaces are then given by

\[
V^\Gamma_{l_1} := \Pi(\Lambda_\gamma(l_1); \mathbb{R}).
\]

By definition, the index sets \(\Lambda_\gamma(l_1)\) are nested for a fixed \(\gamma > 0\) and hence also the \(V^\Gamma_{l_1}\) in the sense of (4.16). The detail spaces \(W^\Gamma_{l_1}\) then consist exactly of the multivariate orthogonal polynomials \(P_\nu\) corresponding to the indices in \(\Lambda_\gamma(l_1) \setminus \Lambda_\gamma(l_1 - 1)\), i.e.

\[
W^\Gamma_{l_1} = \Pi(\Lambda_\gamma(l_1) \setminus \Lambda_\gamma(l_1 - 1); \mathbb{R}).
\]
7.1.3 Approximation results

Define the $L^2_\rho$ projection $P_{l_1}^\Gamma : L^2_\rho(\Gamma) \rightarrow V_{l_1}^\Gamma$ by $P_{l_1}^\Gamma u_M := u_{M,\Lambda_\gamma(l_1)}$ as in (7.11). The following approximation properties rely on the decay of the Karhunen-Loève coefficients $\psi_m$ (4.2), i.e. on the parameter $s > 0$, where

$$\|\psi_m\|_{L^\infty(D)} = \sqrt{\lambda_m} \|\phi_m\|_{L^\infty(D)} \leq C_m^{-s}.$$  

Proposition 7.1.2. Let $s$ be the decay rate of $\|\psi_m\|_{L^\infty(D)}$. If $u_M$ solves (4.5), then for each $0 < r < s - \frac{3}{2}$ exists a constant $C(r)$ such that for every $\gamma > 0$ and for the sequence of projections $P_{l_1}^\Gamma$ corresponding to the index sets $\Lambda_\gamma(l_1)$ in (7.16) it holds

$$\|u_M - P_{l_1}^\Gamma u_M\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq C(r) N_{l_1}^\Gamma \|u_M\|_{A_r(H^1_0(D))}$$  

(7.18)

where $N_{l_1}^\Gamma := |\Lambda_\gamma(l_1)| \rightarrow \infty$ as $l_1 = 0, 1, 2, ...$. Since $N_{l_1}^\Gamma = [2^{\gamma l_1}]$, we can express (7.18) also as

$$\|u_M - P_{l_1}^\Gamma u_M\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq C(r) 2^{-\gamma l_1} \|u_M\|_{A_r(H^1_0(D))}$$  

(7.19)

Here, the constant $C(r)$ is independent of the KL truncation order $M$, which equals the dimension of the stochastic parameter space. The proof will be provided at the end of this chapter.

Remark 7.1.3. To have $r > 0$ in (7.18) requires $s > \frac{3}{2}$ in (2.25). Such a decay can be expected, for example, if $V_a \in H^{t,t}(D \times D)$ with $t > 4d$ as it can be easily derived from Corollary 2.2.8. In other words, the regularity of the 2-point covariance $V_a$ (2.2) implies the range of possible rates $r$ of best $N$-term approximations of $u_M$.

Remark 7.1.4. Proposition 7.1.2 indicates only the existence and the convergence rate $r$ of a semidiscrete best-$N^\Gamma$-term approximation. These approximations are not constructive: the proof uses the (a-priori unknown) values of $\|u_\nu\|_{H^1_0(D)}$. However, in Section 8.1 we will present a strategy based on the Karhunen-Loève eigenvalues of the input data to a-priori locate sequences of sets $\Lambda_\gamma(l_1)$ that appear to be close to the optimal sets in numerical experiments. Moreover, the algorithm to find those sets scales linearly (up to logarithms) in time and memory w.r.t. the number of elements $|\Lambda_\gamma(l_1)|$.

7.2 Analysis of the sparse tensor stochastic Galerkin method

After having separately discussed the convergence rates of the polynomial chaos approximation in $\Gamma$ and the multilevel discretizations in $D$ of our model problem (4.4), we will
now derive approximation results for the sparse tensor stochastic Galerkin formulation (4.25).

To obtain estimates on the convergence rate for the sGFEM approximation \( \hat{u}_M \) of \( u \) it remains, due to the quasi-optimality (4.26), to estimate the best-approximation error of \( u_M \) from the sparse tensor product space \( V^\Gamma_L \otimes V^D_L \).

To this end, we define the sparse tensor product projection operator \( \hat{P}_L : L^2_\rho(\Gamma) \otimes H^1_0(D) \longrightarrow V^\Gamma_L \otimes V^D_L \) by

\[
(\hat{P}_L v)(y, x) := \sum_{0 \leq l_1 + l_2 \leq L} (P^\Gamma_{l_1} - P^\Gamma_{l_1-1}) \otimes (P^D_{l_2} - P^D_{l_2-1}) v(y, x) \quad (7.20)
\]

To simplify the notation, we denote by \( \| \cdot \| \) the norm taken w.r.t. \( L^2_\rho(\Gamma; H^1_0(D)) \) unless otherwise indicated. We denote by \( u_M \) the solution to the problem (4.5) and by \( \hat{u}_M \) the discrete solution to (4.25). To estimate the discretization error \( e_L := \| u_M - \hat{P}_L u_M \| \), we start from the definition of \( \hat{P}_L \) in (7.20) and write

\[
e_L \leq \left| \sum_{l_1 = 0}^{L} \sum_{l_2 = L-l_1+1}^{\infty} (P^\Gamma_{l_1} - P^\Gamma_{l_1-1}) \otimes (P^D_{l_2} - P^D_{l_2-1}) u_M \right|
\]

\[
+ \left| \sum_{l_1 = L+1}^{\infty} \sum_{l_2 = 0}^{\infty} (P^\Gamma_{l_1} - P^\Gamma_{l_1-1}) \otimes (P^D_{l_2} - P^D_{l_2-1}) u_M \right|
\]

\[
= \left| \sum_{l_1 = 0}^{L} (P^\Gamma_{l_1} - \text{Id} + \text{Id} - P^\Gamma_{l_1-1}) \otimes (\text{Id} - P^D_{L-l_1+1}) u_M \right|
\]

\[
+ \left| (\text{Id} - P^\Gamma_{L}) \otimes (\text{Id}) u_M \right|
\]

\[
\leq \sum_{l_1 = 0}^{L} \left( \| (\text{Id} - P^\Gamma_{l_1}) \otimes (\text{Id} - P^D_{L-l_1+1}) u_M \| \right.
\]

\[
+ \left. \| (\text{Id} - P^\Gamma_{l_1-1}) \otimes (\text{Id} - P^D_{L-l_1+1}) u_M \| \right)
\]

\[
+ C(r)2^{-L\gamma r} \| u_M \|_{A_r(H^{1+\gamma} \cap H^1_0(D))}
\]

\[
\leq (C(r) L 2^{-L \min(\gamma r, t)} + C(r) 2^{-L \gamma r}) \| u_M \|_{A_r(H^{1+\gamma} \cap H^1_0(D))}.
\]

Note that we used the approximation results (7.19) and (5.6) in the last two lines of the estimates above. We equilibrate the two error bounds by choosing

\[
\gamma := \frac{t}{r} \quad (7.21)
\]
and we obtain for the error $e_L$ the estimate
\[
    e_L = \|u_M - \hat{u}_M\|_{L_2^p(\Gamma; H_0^1(D))} \leq C(r, t) L^{2-Lt} \|u_M\|_{A_r(H^{1+t} \cap H_0^1(D))} \quad (7.22)
\]
with a constant possibly depending on the stochastic approximation rate $r$.

Since the cardinality of the detail spaces $W_{l_1}^\Gamma, W_{l_2}^D$ both satisfy the assumptions of Lemma 4.2.1, setting $b_\Gamma = 2^\gamma$ and $b_D = 2^d$, we arrive at our main result on convergence rates for the sGFEM discretization of sPDEs.

**Proposition 7.2.1.** Let the solution $u_M$ to the model problem (4.5) satisfy
\[
    u_M \in A_r((H^{1+t} \cap H_0^1(D)) \quad \text{for some } 0 < r < s - \frac{3}{2}, \quad 0 < t \leq p \quad (7.23)
\]
with $s > \frac{3}{2}$ as in (2.25). Let $\hat{u}_M$ denote the sparse tensor approximation to the problem (4.25) with the sparse tensor product spaces sequence $V_{l_1}^\Gamma \hat{\otimes} V_{l_2}^D$, defined in (4.24).

Then, there exists a constant $C > 0$, independent of $M$ and $L$, such that
\[
    \|u_M - \hat{u}_M\|_{L_2^p(\Gamma; H_0^1(D))} \leq C L^{1+\beta} (\hat{N}_L)^{-\beta} \|u_M\|_{A_r((H^{1+t} \cap H_0^1(D))} \quad (7.24)
\]
where $\hat{N}_L := \dim(V_{l_1}^\Gamma \hat{\otimes} V_{l_2}^D)$ and $\beta = \min(r, t/d)$.

We can see from the above definition of $\beta$ that by choosing $p = \lceil rd \rceil$ as the polynomial degree in the spatial approximation space (5.2), we can obtain the maximum possible rate $r$, provided that $u_M$ is sufficiently regular.

**Remark 7.2.2 (full tensor product).** The main motivation for using the sparse tensor product (4.24) between hierarchic sequences of spatial and stochastic discretization spaces lies in the reduction of degrees of freedom to $O((\max(\dim(V_{l_1}^\Gamma), \dim(V_{l_2}^D))))$, shown in Lemma 4.2.1, as opposed to $O(\dim(V_{l_1}^\Gamma) \times \dim(V_{l_2}^D))$ in a full tensor approach. Using the approximation results (7.19) and (5.6), one can derive the full tensor error estimate
\[
    \|u_M - P_L^\Gamma \otimes P_L^D u_M\| \leq \|(\text{Id} - P_L^\Gamma) \otimes (\text{Id} - P_L^D) u_M\| + \|(\text{Id} - P_L^\Gamma) \otimes \text{Id} u_M\|
    + \|(\text{Id} \otimes (\text{Id} - P_L^D) u_M\|
    \leq (C(r, t) 2^{-2L \min(\gamma r, t)} + C(t) 2^{-Lt}) + C(r) 2^{-L \gamma r} \|u_M\|_{A_r((H_0^1 \cap H^{1+t})(D))}.
\]
Choosing, as in (7.21), $\gamma = t/r$, yields
\[
    \|u_M - P_L^\Gamma \otimes P_L^D u_M\| \leq C(r, t) 2^{-Lt} \|u_M\|_{A_r((H_0^1 \cap H^{1+t})(D))}, \quad t \in [0, p].
\]
Since the total number of degrees of freedom in the full tensor case equals $N_L = 2^{(\gamma + d)L} = 2^{(t/r + d)L}$, we obtain the full tensor estimate

$$
\|u_M - P_L^D \otimes P_L^F u_M\| \leq C(\bar{N}_L)^{-\bar{\beta}}\|u\|_{A((H^{t\wedge 1}_0 \cap H^{t+1}(D)))}, \quad t \in [0, p].
$$

(7.25)

with rate $\bar{\beta} = (d/t + 1/r)^{-1}$ and where $P_L^D$ and $P_L^F$ denote the $H^1_0$- and $L^2$-projections onto the spatial and stochastic discretization spaces, respectively.

A proof of the regularity (7.23) for $p = 1$ is given in [12].

### 7.3 Proof of Proposition 7.1.2

To prove Proposition 7.1.2, we will have to provide a bound on the coefficients $u_\nu(x)$ in (7.8) in both, the Legendre and the Hermite case. To this end, we will essentially use the analyticity property of the solution, cf. Assumption 3.2.8, and therefore first extend the parametric formulation of our problem (4.36) to complex valued arguments. We then establish some results on approximation of sequences of such coefficients and complete the section by the actual proof of Proposition 7.1.2.

#### 7.3.1 Analytic continuations

Recall the definition of $\Sigma(\Gamma_m, \tau_m)$ from (3.9) and define the Cartesian product domain

$$
\Sigma(\Gamma, \tau) = \times_{m=1}^M \Sigma(\Gamma_m, \tau_m) \subset \mathbb{C}^M,
$$

(7.26)

where $\tau = (\tau_1, \ldots, \tau_M) \in \mathbb{R}^M$.

We adopt the notation from Section 3.2.5 and write $y_m^* = (y_1, \ldots, y_m-1, y_{m+1}, ...)$ and $y = (y_m^*, y_m) \in \Gamma_m^* \times \Gamma_m$ if the dependence of $y$-dependent quantities on the coordinate $y_m$ is to be emphasized. Likewise, we denote by $z = (z_m^*, z_m) \in \Sigma(\Gamma, \tau) = \Sigma(\Gamma_m^*, \tau_m^*) \times \Sigma(\Gamma_m, \tau_m)$ the corresponding partitioning of the complex vector $z$. Recall the definition of the real-valued parametric bilinear form $b_M(y, \cdot, \cdot)$ from (4.37), which we will next extend to complex-valued arguments: abusing notation, we denote by $V^D = H^1_0(D)$ also the space of complex-valued functions which belong to $H^1_0(D)$ and define, for $z \in \Sigma(\Gamma, \tau)$ the complex extension of the (truncated) diffusion coefficient $a_M$ (2.11),(2.28) by either

$$
a_M(z, x) = E_a(x) + \sum_{m=1}^M \sqrt{\lambda_m} \phi_m(x) z_m
$$

(7.27)
in the case of bounded domains $\Gamma_m$, or by

$$a_M(z, x) = e^{\log(x) + \sum_{m=1}^{\infty} \sqrt{\lambda_m} \varphi_m(x) z_m}$$

(7.28)

in the case of unbounded domains. For $u, v \in V^D$ the sesquilinear extension of $B_M(y; \cdot, \cdot)$ is then given by

$$B_M(z; u, v) = \int_D a_M(z, x) \nabla u \cdot \nabla v \, dx.$$  

(7.29)

The complex valued parametric problem is then given by

$$\begin{cases} 
-\text{div}(a_M(z, x) \nabla u_M(z, x)) = f(x) \quad \text{in } D, \\
|u_M(z, x)|_{x \in \partial D} = 0.
\end{cases}$$

(7.30)

and, in weak form: Find $u_M \in L^2(\Sigma(\Gamma, r); H^1_0(D))$ s.t.

$$B_M(z; u_M, v) = \int_D f v \, dx \quad \forall v \in H^1_0(D) \quad \forall z \in \Sigma(\Gamma, r).$$

(7.31)

We will now prove that for certain $r = (r_1, \ldots, r_M)$, problem (4.5) with $b_M$ as in (7.31) is uniquely solvable in $\Sigma(\Gamma, r)$ and its solution is ‘coordinate-wise’ analytic. We start with the case of bounded domains $\Gamma_m$.

**Lemma 7.3.1.** Let $a_M$ be given as in (7.27). Define the vector of radii $r(\delta)$ by

$$r_m(\delta) := \frac{\tau_m}{C(\delta)m^{1+\delta}}, \quad \text{where} \quad 0 < \tau_m < \frac{a^*_\text{min}}{\sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)}}$$

(7.32)

and

$$C(\delta) := \sum_{m=1}^{\infty} \frac{1}{m^{1+\delta}}, \quad a^*_\text{min} = \min_{x \in D} \mathbb{E}_a(x) - a_*,$$

(7.33)

with $a_*$ given in (3.13). Then, for any $z \in \Sigma(\Gamma, r(\delta))$ with $\delta > 0$, (4.5) admits a unique solution $u_M \in L^2(\Sigma(\Gamma, r(\delta)); H^1_0(D))$. Moreover, $u_M(z^*_m, z_m; \cdot) : \Sigma(\Gamma, r(\delta)) \to H^1_0(D)$ is, for fixed coordinates $z^*_m \in \Sigma(\Gamma^*_m, r^*_m(\delta))$, a $H^1_0(D)$-valued analytic function of $z_m \in \Sigma(\Gamma_m, r_m(\delta))$

**Proof.** To prove the existence and uniqueness of a solution to (7.31), we show that the real part $\text{Re} a(z, x)$ of the random field, as in (7.27), with $z \in \Sigma(\Gamma, r(\delta))$, is bounded away from zero:

$$\text{Re} a_M(z, x) = \mathbb{E}_a(x) + \sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x) \text{Re} z_m$$

$$\geq \mathbb{E}_a(x) - \sum_{m \geq 1} b_m |z_m|$$

(7.34)
\[ \geq \mathbb{E}_a(x) - \sum_{m \geq 1} b_m \frac{a_{\min}^*}{C(\delta) b_m m^{1+\delta}} \]
\[ \geq \min_{x \in D} \mathbb{E}_a(x) - a_{\min} =: a_{\min}^* > 0, \]

where \( b_m := \| \psi_m \|_{L^\infty(D)} \). In a similar fashion we can conclude that
\[ |a_M(z, \mathbf{x})| \leq a_{\max}^* := \| \mathbb{E}_a \|_{L^\infty(D)} + a_{\min}^* \]
for \( \mathbf{z} \in \Sigma(\Gamma, r(\delta)) \) and \( \mathbf{x} \in D \). It follows that \( \forall \mathbf{z} \in \Sigma(\Gamma, r(\delta)) \) and \( \forall u, v \in H^1_0(D) \)
\[ \text{Re} B(z; u, u) \geq \frac{1}{a_{\min}^*} \| u \|^2_{H_0^1(D)} \quad \text{and} \quad |B(z; u, v)| \leq a_{\max}^* \| u \|_{H^1_0} \| v \|_{H^1_0}. \]

(7.34)

Hence, the unique solvability of (4.5) for all \( \mathbf{z} \in \Sigma(\Gamma, r(\delta)) \) follows by Lax-Milgram. Choosing \( v = u_M(z, \cdot) \) in (7.31), we find with (7.34) that
\[ a_{\min}^* \| u_M(z, \cdot) \|^2_{H^1_0} \leq |B(z; u_M(z, \cdot), u_M(z, \cdot))| = |(f, u_M(z, \cdot))| \leq \| f \|_{H^{-1}} \| u_M(z, \cdot) \|_{H^1_0}, \]
from where it follows that
\[ \forall \mathbf{z} \in \Sigma(\Gamma, r(\delta)) : \| u_M(z, \cdot) \|_{H^1_0} \leq \frac{\| f \|_{H^{-1}}}{a_{\min}^*} \]
(7.35)

and hence in particular that \( u_M \in L^2_0(\Gamma; H^1_0(D)) \). We now establish analyticity with respect to \( \mathbf{z}_m \) by a power series argument. The proof can either be done analogous to the one for Proposition 3.3.3 or, due to the linear dependence of \( a_M \) on \( y_m \), also more directly. For reasons of completeness we will consider the latter option. Hence, we consider the parametric problem (7.30) for parameters \( (\mathbf{z}_m, y_m) \in \Sigma(\Gamma_m, r_m(\delta)) \times \Gamma_m \). Differentiating (7.30) with respect to the real-valued parameter \( y_m \), we obtain that for any \( \nu \in \mathbb{N}, \mathbf{z}_m \in \Sigma(\Gamma_m, r_m(\delta)) \) and every \( v \in H^1_0(D) \), the partial derivative \( \partial_{y_m} u_M((\mathbf{z}_m, y_m); \cdot) \in H^1_0(D) \) solves the problem:

\[ B((\mathbf{z}_m, y_m); \partial_{y_m} u, v) = -\nu \int_D \sqrt{\lambda_m} \varphi_m(x) \nabla (\partial_{y_m}^{-1} u_M) \cdot \nabla v \, dx \quad \forall v \in H^1_0(D). \]

Choosing here \( v = \partial_{y_m} u_M(z_m, y_m, \cdot) \in H^1_0(D) \) gives
\[ a_{\min}^* \| \partial_{y_m} u_M(z_m, y_m, \cdot) \|^2_{H^1_0} \leq |B((\mathbf{z}_m, y_m); \partial_{y_m} u_M, \partial_{y_m} u_M)| \]
\[ = \nu \left| \int_D \sqrt{\lambda_m} \varphi_m(x) \nabla (\partial_{y_m}^{-1} u_M(z_m, y_m, x) \times \nabla \partial_{y_m} u_M(z_m, y_m, x) \, dx \right| \]

70
7.3 Proof of Proposition 7.1.2

\[ \leq \nu \sqrt{\lambda_m} \| \varphi_m \|_{L^\infty(D)} \| \partial_{y_m}^{\nu-1} u_M(z_m^*, y_m, \cdot) \|_{H^1_0} \]

\[ \times \| \partial_{y_m}^\nu u_M(z_m^*, y_m, \cdot) \|_{H^1_0}, \]

whence we obtain, for any \( z_m^* \in \Sigma(\Gamma_m^*, r_m^*(\delta)) \) and any \( m, \nu \in \mathbb{N} \),

\[ \| \partial_{y_m}^\nu u_M(z_m^*, y_m, \cdot) \|_{H^1_0} \leq \nu \frac{b_m}{a_{\min}} \| \partial_{y_m}^{\nu-1} u_M(z_m^*, y_m, \cdot) \|_{H^1_0}. \]

Iterating this estimate gives with (7.35)

\[ \| \partial_{y_m}^\nu u_M(z_m^*, y_m, \cdot) \|_{H^1_0} \leq \nu! \left( \frac{b_m}{a_{\min}} \right)^\nu \| u_M(z_m^*, y_m, \cdot) \|_{H^1_0} \]

\[ \leq \nu! \left( \frac{b_m}{a_{\min}} \right)^\nu \| f \|_{H^{-1}}. \quad (7.36) \]

Now define, for arbitrary \( m \in \mathbb{N} \) and arbitrary, but fixed \( y_m \in \Gamma_m \), the formal power series \( \Sigma(\Gamma_m, r_m(\delta)) \ni z_m \rightarrow u_M(z_m^*, z_m, \cdot) \in H^1_0(D) \) by

\[ u_M(z_m^*, z_m, \cdot) = \sum_{k=0}^{\infty} \frac{(z_m - y_m)^k}{k!} \partial_{y_m}^k u_M(z_m^*, y_m, \cdot). \quad (7.37) \]

To estimate its convergence radius, we use (7.36):

\[ \| u_M(z_m^*, z_m, \cdot) \|_{H^1_0} \leq \sum_{k=0}^{\infty} \frac{|z_m - y_m|^k}{k!} \| \partial_{y_m}^k u_M(z_m^*, y_m, \cdot) \|_{H^1_0} \]

\[ \leq \| f \|_{H^{-1}} \sum_{k=0}^{\infty} \left( \frac{b_m |z_m - y_m|}{a_{\min}} \right)^k. \]

This infinite sum, and hence the series (7.37), converges in \( H^1_0(D) \) if

\[ b_m \frac{|z_m - y_m|}{a_{\min}} < 1. \]

Since \( y_m \in \Gamma_m \) was arbitrary, the series (7.37) converges for any \( y_m \in \Gamma_m = [-1, 1] \) and, by a continuation argument, for any \( z_m \in \mathbb{C} \) satisfying \( \text{dist}(z_m, \Gamma_m) < a_{\min}/b_m \). This completes the proof.

We now prove an analogous result in the case of lognormal coefficients (7.28).
Lemma 7.3.2. Let $a_M$ be given as in (7.28). Define the vector of radii $r(\delta)$ by

$$r_m(\delta) := \frac{\pi \tau_m}{2C(\delta)m^{1+\delta}}, \quad \text{where} \quad 0 < \tau_m < \frac{1}{2\sqrt{\lambda_m} \varphi_m \|L^\infty(D)}$$

and

$$C(\delta) := \sum_{m \geq 1} \frac{1}{m^{1+\delta}}.$$

Then, for any $\delta > 0$, (4.5) admits a unique solution $u_M \in L^2_G(\Sigma(\Gamma, r(\delta)), H^1_0(D))$, where $G$ is the standard Gauss measure on $\mathbb{R}$ as constructed in Section 2.2.4. Moreover, the solution $u_M(z^*_m, z_m, \cdot) : \Sigma(\Gamma, r^*_m(\delta)) \rightarrow H^1_0(D)$ is a $H^1_0(D)$-valued analytic function of $z_m \in \Sigma(\Gamma, r_m(\delta))$.

Proof. In the case of unbounded domains, we have to show that the real part $\text{Re} a_M$ of the random field given in (7.28) can be bounded from below by some function $a_*(z) > 0$. The existence and uniqueness of a solution to (7.31) in $L^2_G(\Sigma(\Gamma, r(\delta)), H^1_0(D))$ then follows essentially as in Remark 3.3.1. Denoting $z_m = y_m + iw_m$ we have

$$a_M(z, x) = \exp \left( \sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x)y_m \right) \exp \left( \sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x)w_m \right).$$

It follows

$$\text{Re} a_M(z, x) = \exp \left( \sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x)y_m \cos \left( \sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x)w_m \right) \right).$$

Hence, to ensure that the real part of $a_M$ is positive, we require

$$\sum_{m \geq 1} \sqrt{\lambda_m} \varphi_m(x)w_m \leq b_m r_m(\delta) < \frac{\pi}{2},$$

which follows from (7.38). A lower bound for $\text{Re} a_M$ is then given, as in (3.22), by

$$a_*(z) = \exp \left( \min \text{Re} a_M(z, x) - \sum_{m \geq 1} \|\psi_m\|_{L^\infty(D)} |\text{Re} z_m| \right),$$

where $z_m = y_m + iw_m$, $|w_m| \leq r_m(\delta)$ and existence and uniqueness follow by an application of the Lax-Milgram Lemma as in Remark 3.3.1. The proof of analyticity then is analogous to the one given for Proposition 3.3.3. \qed
Now we can prove analyticity of $\Sigma(\Gamma, r(\delta)) \ni z \rightarrow u(z, \cdot) \in H^1_0(D)$.

**Theorem 7.3.3.** Given the assumptions of Lemma 7.3.1 and 7.3.2, respectively, for any $M \in \mathbb{N}$, the solution $u_M$ to (4.5) is, as a function of $z$, analytic in the complex domain $\Sigma(\Gamma, r(\delta))$ with $r(\delta)$ defined as in (7.32) or (7.38), respectively.

**Proof.** This is a direct consequence of Hartogs’ Theorem, see e.g. [32], Theorem 2.2.8, stating that coordinate-wise analyticity implies global analyticity. 

This enables us in the following sections to find bounds on the Legendre and Hermite ‘coefficients’ $u_\nu \in H^1_0(D)$ due to Cauchy’s integral formula for analytic functions of several variables ([32], Theorem 2.2.1).

**7.3.2 Bounds on the Legendre coefficients**

**Lemma 7.3.4.** If $u_M$ solves the parametric deterministic problem (4.5) and is expanded in a Legendre chaos series as in (7.8) with $\mathcal{P}_\nu = \mathcal{L}_\nu$, then, for any $\delta_1 > 0$, the Legendre coefficients satisfy

$$
\|u_\nu\|_{H^1_0(D)} \leq C \left( \prod_{m \in \text{supp}(\nu)} 2(2\nu_m + 1) \right) \eta(\delta_1)^{-\nu} \|u_M\|_{L^\infty(\Sigma(\Gamma, r); H^1_0(D))},
$$

(7.39)

where $C > 0$ is independent of $M, \nu$ with $\eta(\delta_1) = (\eta_1(\delta_1), \ldots, \eta_M(\delta_1)) \in \mathbb{R}^M$ and

$$
\eta_m(\delta) := r_m(\delta) + \sqrt{1 + r_m(\delta)^2}
$$

(7.40)

and $r_m(\delta)$ as in (7.32).

**Proof.** Note that for notational convenience we do not indicate in the following the dependence of $\eta$ on $\delta_1$. The $H^1_0(D)$-valued Legendre ‘coefficients’ are given by

$$
u \nu = \left( \prod_{m=1}^M 2\nu_m + 1 \right) \int_\Gamma u_M(y) \mathcal{L}_\nu(y) \rho(y) \, dy,
$$

(7.41)

where the integral is a Bochner integral w.r.t. the probability measure $d\rho(y) = \rho(y)dy$ and the equality has to be read in $H^1_0(D)$.

In the following, to keep the results general, we will denote by $W(D)$ an arbitrary Banach space. Moreover, we will use the abbreviations $S = \text{supp}(\nu) \subset \{1, \ldots, M\}$ and $\overline{S} = \{1, \ldots, M\} \setminus S$ (omitting the dependence on $\nu$ for reasons of readability of the formulae below) and, for a set $G \subset \{1, \ldots, M\}$, we will write $\Gamma_G = \prod_{m \in G} \Gamma_m$.
Using Bochner’s Theorem 1.3.1 together with (7.45) and (7.43), we obtain

\[ \eta_m = r_m(\delta_1) + \sqrt{1 + r_m(\delta_1)^2} > 1, \]

(7.42)

to rewrite the Legendre coefficients (7.41) as

\[
u = \left( \prod_{m \in S} \frac{2\nu_m + 1}{2\pi i} \right) \int_{\Gamma} \mathcal{L}_\nu(y_S) \int_{\mathcal{E}_S} \frac{u_M(z_S, y_S)}{(z_S - y_S)} dz_S \rho(y) dy
\]

(7.43)

By [58] §4.9, the innermost integral in (7.43) is a representation for the Legendre polynomials of the second kind,

\[ Q_\nu(z_S) = \int_{\Gamma_S} \frac{\mathcal{L}_\nu(y_S)}{(z_S - y_S)} \rho(y) dy_S. \]

(7.44)

Furthermore, if, for \( m \in S \), we substitute \( z_m = \frac{1}{2}(w_m + w_m^{-1}) \) (Joukowski transformation) with \( |w_m| = \eta_m \), the Legendre polynomials of the second kind can be expanded like (cf. [16], Lemma 12.4.6)

\[ Q_{\nu_m}(\frac{1}{2}(w_m + w_m^{-1})) = \sum_{k=\nu_m+1}^{\infty} \frac{q_{\nu_m k}}{w_m^k} \]

with \( |q_{\nu_m k}| \leq \pi \). Hence,

\[ |Q_\nu(z_S)| \leq \prod_{m \in S} \sum_{k=\nu_m+1}^{\infty} \frac{\pi}{\eta_m^k} = \prod_{m \in S} \frac{\pi^{1-\nu_m}}{1-\eta_m}. \]

(7.45)

Using Bochner’s Theorem 1.3.1 together with (7.45) and (7.43), we obtain

\[
\|u_\nu\|_{W(D)} \leq \left( \prod_{m \in S} \frac{2\nu_m + 1}{2\pi i} \right) \int_{\Gamma} \mathcal{L}_\nu(y_S) |Q_\nu(z_S)| dz_S \rho_S(y_S) dy_S
\]

\[
\leq \left( \prod_{m \in S} \frac{2\nu_m + 1}{2\pi i} \right) \int_{\Gamma_S} \int_{\mathcal{E}_S} \|u_M(z_S, y_S)\|_{W(D)} Q_\nu(z_S) dz_S \rho_S(y_S) dy_S
\]

\[
\leq \left( \prod_{m \in S} \frac{2\nu_m + 1}{2\pi} \operatorname{Len}(E_m) \right) \|u_M(z)\|_{L^1(\mathcal{E}_S \times \Gamma_S)} \max_{z_S \in \mathcal{E}_S} |Q_\nu|.
\]
7.3 Proof of Proposition 7.1.2

\[ \leq \left( \prod_{m \in S} \frac{2\nu_m + 1}{2\pi} \text{Len}(E_m) \right) \prod_{m \in S} \frac{\eta_m - \nu_m - 1}{1 - \eta_m} \| u_M(z) \|_{L^\infty(E_S \times \Gamma_S \mathcal{W}(D))} \]

\[ \leq C \left( \prod_{m \in S} 2(2\nu_m + 1) \right) \eta^{-\nu} \| u_M(z) \|_{L^\infty(E_S \times \Gamma_S \mathcal{W}(D))} \]

(7.46)

where the last estimate holds due to the fact that \( \text{Len}(E_m) \leq 4\eta_m \). The constant is given by \( C = \prod_{m=1}^{M} \frac{1}{1 - \eta_m} \).

\[ \square \]

Remark 7.3.5. Note that the constant \( C \) is bounded independently of \( M \), since the \( \eta_m \)'s are summable and hence the product remains finite even if \( M \to \infty \).

Under the assumption that the covariance \( V_\alpha \in H^{t,d}(D \times D) \) with \( t > d/2 \), it follows from Corollary 2.2.8 that for \( \delta_1 > 0 \) sufficiently small there exists a constant \( C(s,d,\delta_1) > 0 \) such that

\[ \forall m \in \mathbb{N} : \quad \eta_m^{1} \leq C_1(s,d,\delta_1)m^{-(s-1-\delta_1)} \]

(7.47)

where \( s \) is defined in (2.26). Since, furthermore, for any \( \delta_2 > 0 \) there exists a constant \( C_2(\delta_2) \) s.t. \( 2\nu_m + 1 \leq C_2(\delta_2)\eta_m^{2\delta_2} \), it follows from the definition of \( \eta_m \) together with (7.47)

Corollary 7.3.6. Given the assumptions of Lemma 7.3.4. For any \( 0 < t < s - 1 \) there exists a constant \( C_3(t,d) > 0 \), independent of \( m,M,\nu \) s.t. by defining

\[ \tilde{\eta}_m^{-\nu_m} := \begin{cases} 1 & \nu_m = 0 \\ C_3(t,d)\eta_m^{-(1-\delta_2)\nu_m} & \nu_m > 0 \end{cases} \]

(7.48)

the Legendre coefficients satisfy

\[ \| u_\nu \|_{H^s_0(D)} \leq C \prod_{m \in \text{supp}(\nu)} \tilde{\eta}_m^{-\nu_m} \]

(7.49)

where the constant \( C > 0 \) is as in Lemma 7.3.4 and, by Remark 7.3.5, does not depend on \( M,\nu \).

Note that it follows from (7.47), that \( \tilde{\eta}_m^{-\nu_m} \to 0 \) if either \( m \) or \( \nu_m \) tend to infinity, hence the product in (7.49) remains finite even if \( |\text{supp}(\nu)| \to \infty \).

7.3.3 Bounds on the Hermite coefficients

Lemma 7.3.7. If \( u_M \in C^0(\Gamma, H^1_0(D)) \) solves the parametric deterministic problem (4.5) and is expanded in a Hermite chaos series as in (7.8) with \( \mathcal{P}_\nu = H_\nu \), then, for any \( \delta_1 > 0 \) the Hermite coefficients satisfy

\[ \| u_\nu \|_{H^s_0(D)} \leq \tilde{C}C(\alpha)\eta(\delta_1)^{-\nu} \| u_M \|_{C^0(\Sigma,H^s_0)}, \]

(7.50)
where \( C > 0 \) is independent of \( M, \nu, \alpha \),

\[
C(\alpha) = \prod_{m=1}^{M} e^{\frac{\alpha_m^2}{2}} \left( 1 + \text{Erf} \left( \frac{\alpha_m}{\sqrt{2}} \right) \right),
\]

(7.51)

and \( \alpha_m \) given by (3.7) with

\[
\eta(\delta) := r(\delta)
\]

(7.52)

and \( r_m(\delta) \) as in (7.38).

Remark 7.3.8. From Section 3.3.3, together with (3.42) and Corollary 2.2.8, we have that \( \alpha_m \sim m^{-\delta} \). Hence, if \( s > \frac{1}{2} \) (entailing \( V_a \in H^t(D \times D) \) with \( t > 2d \)) then the product in (7.51) is finite, independent of \( M \).

**Proof of Lemma 7.3.7.** We will use the same notations and abbreviations as in the proof of 7.3.4. The \( H^0_\nu \)-valued Hermite ‘coefficients’ are given by

\[
u(\nu) = \frac{1}{\nu!} \int_{\Gamma} u_M(y) H_\nu(y) \rho(y) \, dy,
\]

(7.53)

where \( \rho_m \) denotes the standard Gauss probability density function \( \rho_m(y_m) = \frac{1}{\sqrt{2\pi}} e^{-y_m^2/2} \).

By \( C_m \subset \Sigma(\mathbb{R}, \eta_m) \), we denote the circle centered at the origin with radius \( \eta_m \) and by \( C_S = \prod_{m \in \text{supp}(\nu)} C_m \). Using Rodrigues’ formula (6.34) together with integration by parts and Cauchy’s integral formula, we can rewrite the Hermite coefficients (7.53) as

\[
u(\nu) = \frac{(-1)^{|\nu|}}{\nu!} \int_{\Gamma} u_M(y) \partial^{\nu} \rho(y) \, dy = \frac{1}{\nu!} \int_{\Gamma} (\partial^{\nu} u_M(y)) \rho(y) \, dy
\]

(7.54)

By Bochner’s Theorem 1.3.1 we then obtain

\[
\| u_\nu \|_{W(D)} \leq \frac{\| u \|_{C_0^\infty(\Sigma \times \Gamma ; H^0_\nu)}}{(2\pi)^{|\nu|_0}} \int_{\Sigma} \int_{\Gamma} \prod_{m \in S} e^{\alpha_m |\text{Re} z_m|} \frac{1}{y_m^{\nu+1}} \, dz \, dy
\]

\[
\times \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi}} e^{-y_m^2/2+\alpha_m |y_m|} \, dy
\]

Since \( \sup_{z_m \in C_m} |\text{Re} z_m| \leq \eta_m \) and

\[
\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1}{2} y_m^2 + \alpha_m |y_m|} \, dy_m = e^{\frac{1}{2} \alpha_m} \left( 1 + \text{Erf} \left( \frac{\alpha_m}{\sqrt{2}} \right) \right),
\]

76
we obtain
\[ \|u_\nu\|_{W(D)} \leq C(\alpha)\eta(\delta_1)^{-\nu} \left( \prod_{m \in S} e^{\alpha_m \eta_m} \right) \|u_M\|_{C_0(\Sigma;W(D))} \]
with \( C(\alpha) \) as in (7.51). From Section 3.3.3 and (3.42) we have that \( \alpha_m \sim \tau^{-1}_m \). On the other hand, by (7.52) and (7.38) we have \( \eta_m \sim \tau_1 m^{1+\delta} \), hence
\[ \prod_{m \in S} e^{\alpha_m \eta_m} \leq \exp(c_1 \sum_{m \in S} \frac{1}{m^{1+\delta}}) \leq \exp(c_1 C(\delta)) := \bar{C} \]
with \( C(\delta) \) as in (7.33). Hence, \( \bar{C} \) is independent of \( |\nu|_0 \) and \( M \). This completes the proof. \( \square \)

Under the assumption that the covariance \( V \) is at least in \( H^{t,t}(D \times D) \) with \( t > 2d \), it follows from Corollary 2.2.8 that for \( \delta_1 > 0 \) sufficiently small there exists a constant \( C(s,d,\delta_1) > 0 \) such that
\[ \forall m \in \mathbb{N} : \quad \eta_m^{-1} \leq C_1(s,d,\delta_1) m^{-(s-1-\delta_1)} \quad (7.55) \]
where \( s \) is defined in (2.26).

### 7.3.4 \( \tau \)-summability of the PC coefficients

To obtain best \( N \)-term approximation rates, suitable regularity of the function under consideration is required [18]. This means in our case that the norms of the Legendre or Hermite coefficients \( u_\nu \) have to be \( \tau \)-summable.

**Lemma 7.3.9.** Let \( V_a \in H^{t,t}(D \times D) \) with \( t > 3d \). If \( u_M \) solves (4.5) and is expanded in a Legendre or Hermite series as in (7.8), then for any \( \tau > \frac{1}{s-1} \), where \( s \) is given by (2.26), then \( \{\|u_\nu\|_{H^1(D)} : \nu \in \mathbb{N}^M_0\} \in \ell^\tau(\mathbb{N}^M_0) \) and there holds:
\[ \{\|u_\nu\|_W(D) : \nu \in \mathbb{N}^M_0\} \|_{\ell^\tau(\mathbb{N}^M_0)} = \sum_{\nu \in \mathbb{N}^M_0} \|u_\nu\|_W(D) \leq C \exp \left( \frac{C_3 \eta^{-1}_1 \|\tau_{\mathbb{N}_0^M}\|_{\ell^\tau(\mathbb{N}_0^M)}}{(1 - \eta_1^{-1})} \right) \quad (7.56) \]
with \( C = C(\tau,d) > 0 \) independent of \( M \), \( C_3 \) as in (7.48) in the Legendre case or \( C_3 = 1 \) in the Hermite case and with \( \tilde{\eta}_m \) defined as in (7.40).

**Proof.** From (7.47) or (7.55), respectively, it follows that \( \|(\eta_1^{-1})_m\|_{\ell^\tau} < \infty \) if \( \tau > \frac{1}{s-1} \).
In the Legendre case, using (7.48) together with the summation formula for geometric series, we obtain for any finite $M < \infty$ that

$$\sum_{\nu \in N_0^M} \|u_{\nu}\|_{H^1_0(D)}^2 \leq C \sum_{m=1}^{M} \sum_{\nu_m=0}^{\infty} \prod_{m=1}^{M} \tilde{\eta}_{m}^{-\nu_{m}} \leq C \prod_{m=1}^{M} \left(1 + \frac{C_3 \eta_m}{1 - \eta_m}\right) \leq C e^{\frac{C_3}{1 - C_3 \eta_1}} \prod_{m=1}^{M} \eta_{m}^{-\tau_m},$$

where we used $1 + x \leq e^x$ if $x \geq 0$. Letting $M \to \infty$ gives (7.56).

In the Hermite case, we obtain by a similar argument, but due to slightly better bounds (7.50)

$$\sum_{\nu \in N_0^M} \|u_{\nu}\|_{H^1_0(D)}^2 \leq C \sum_{m=1}^{M} \sum_{\nu_m=0}^{\infty} \prod_{m=1}^{M} \eta_{m}^{-\nu_{m}} \leq C e^{\frac{1}{1 - \eta_1}} \prod_{m=1}^{M} \eta_{m}^{-\tau_m}.$$

7.3.5 Proof of Proposition 7.1.2

Recall the definitions of $\Pi(\Lambda; \mathcal{W}(D))$, $\sigma_N$ and $\mathcal{A}_r(\mathcal{W}(D))$ from (7.13)-(7.15) with $\mathcal{W}(D)$ denoting a Banach space, e.g. $H^1_0(D)$ in our examples. The weak-$\ell_\tau(N)$ spaces $\ell_\tau^w(N)$ consist of all sequences $(f_k)_{k \in N}$ for which it holds

$$\|f_k\|_{\ell_\tau^w(N)} := \sup_{k \geq 1} k^{1/\tau} f_k^* < \infty \quad (7.57)$$

where $(f_k^*)_k$ denotes the decreasing rearrangement of $(|f_k|)_k$. Since $N_0^M$ is a countable set of indices, there exists a bijection $i : N \to N_0^M$, and we can identify

$$u_k = u_{i(k)}, \quad u_k^* = u_{i(k)}^* \quad \forall k \in N.$$

Where appropriate, we will switch between these two representations of the series $\{u_k\}_{k \in N}$ and $\{u_{\nu}\}_{\nu \in N_0^M}$, respectively. Since $\ell_\tau(N_0^M) \subset \ell_\tau^w(N_0^M)$, this enables us now to characterize the space $\mathcal{A}_r(\mathcal{W}(D))$: In fact, a generalization of a well-known result of nonlinear approximation theory to Bochner spaces (see e.g. [18], Theorem 4) claims that a function $u_M \in L^2(\Gamma; \mathcal{W}(D))$ belongs to $\mathcal{A}_r(\mathcal{W}(D))$, if and only if the $\mathcal{W}(D)$-norm of its polynomial chaos coefficients $u_{\nu}$ in (7.8) form a sequence in $\ell_\tau^w(N_0^M)$ for $\tau := (r + \frac{1}{2})^{-1}$. In fact, by virtue of the bijection $i$, denote by $(u_k^*)_{k \in N}$ the decreasing rearrangement of $(\|u_{\nu}\|_{\mathcal{W}(D)})_{\nu}$. Due to

$$|u_{\mathcal{A}_r(\mathcal{W}(D))}|^2 = \sup_{N \geq 1} N^{2r} \sum_{k \geq N} \|u_k^*\|^2_{\mathcal{W}(D)}.$$
7.3 Proof of Proposition 7.1.2

\begin{align*}
\leq & \left\| u_{\nu} \right\|_{W(D)}^2 \left\| \nu \right\|_{L^\infty(S^d_M)}^2 \sup_{N \geq 1} N^{2r} \sum_{k > N} k^{-2r-1} \\
\leq & \left\| u_{\nu} \right\|_{W(D)}^2 \left\| \nu \right\|_{L^\infty(S^d_M)}^2 \sup_{N \geq 1} \int_N^\infty \frac{N^{2r}}{x^{2r+1}} dx \\
\leq & \left\| u_{\nu} \right\|_{W(D)}^2 \left\| \nu \right\|_{L^\infty(S^d_M)}^2 \frac{1}{2r + 1}
\end{align*}

and

\begin{align*}
(u_{2k}^*)^2 - (u_{2k})^2 & \leq \frac{1}{k} \sum_{l=k+1}^{2k} (u_{l}^*)^2 \leq \frac{1}{k} \sigma_k(u)^2 \leq |u|_{A_r(W(D))}^2 k^{-2r-1} 
\end{align*}

we have that

\[ |u|_{A_r(W(D))} \lesssim 2\tau \left\| u_{\nu} \right\|_{W(D)} \left\| \nu \right\|_{L^\infty(S^d_M)} \lesssim \frac{2}{2r + 1} |u|_{A_r(W(D))} \]

Since Lemma 7.3.9 ensures that the Legendre coefficients of the solution \( u_M \) belong to \( \ell^\infty_r(N^d_M) \) for \( \tau > (s - 1)^{-1} \), it follows that the solution \( u_M \) is in \( A_r(W(D)) \) for \( \tau = (r + 1/2)^{-1} > (s - 1)^{-1} \) and hence for \( 0 < r < s - 3/2 \). Proposition 7.1.2 then follows from the definition of \( A_r(W(D)) \) (7.15).
7 Sparse tensor stochastic Galerkin
8 Implementation and numerical examples

In this chapter, we will discuss some of the issues regarding the implementation of the sparse tensor stochastic Galerkin and stochastic collocation method and provide numerical examples. First, we will summarize the stochastic Galerkin and stochastic collocation algorithms and explain some of the algorithmic technicalities which are non-standard. In particular we will describe the heuristic algorithm to find a quasi best-$N$-term approximation of the solution in the random parameter space. We will then provide numerical examples for the sparse tensor Galerkin and collocation algorithms and verify the theoretical results provided in Chapters 5-7.

8.1 Implementational aspects

8.1.1 Algorithms

Given the truncation level $M$ of the Karhunen-Loève expansion, a discretization level $L$ and known statistics $E_a(x)$ (2.1) and $V_a(x,x')$ (2.2) of the input random field $a$, we propose the following algorithms

Algorithm 8.1.1. (sGFEM algorithm).

1. Compute eigenpairs $(\lambda_m, \varphi_m)$, $m = 1, ..., M$ of the covariance operator $V_a$ (2.5).
2. Compute the wavelet basis $\Psi_L$ in physical domain $D$ up to level $L$ as constructed in Section 5.1.
3. Choose the stochastic discretization parameter $\gamma$ as in (7.21) and determine the index set $\Lambda_\gamma(L)$ of the largest PC coefficients $u_\nu$.
   
   Compute the Galerkin stiffness matrix and load vector w.r.t. the sparse tensor product space $V_L^T \otimes V_L^D$.
4. Solve the resulting system of linear equations by a preconditioned CG-algorithm.
5. Calculate the required statistical data from the solution (postprocessing).

Algorithm 8.1.2. (sCFEM algorithm).
1. **Compute** eigenpairs \((\lambda_m, \varphi_m), m = 1, \ldots, M\) of the covariance operator \(\mathcal{V}_a(2.5)\).

2. **Compute** the wavelet basis \(\Psi_L\) up to level \(L\) as constructed in Section 5.1.

3. **Choose** the stochastic discretization parameter \(\gamma\) as in (6.29) and **compute** the Smolyak collocation points \(y_j\) based on the \(\lceil \gamma L \rceil\) Gauss nodes in each stochastic dimension.

4. **Solve** (4.37) for each collocation point \(y_j\) by finite element projection onto the corresponding detail space \(W_{t_2}^{D_2}(4.43)\).

5. **Calculate** the required statistical data from the set of solutions (postprocessing).

### 8.1.2 KL-eigenpair computation

For special covariances, like exponential ones, and simple domains \(D\), closed-form solutions to (2.6) can be obtained by separation of variables, [27]. For more general correlation functions and arbitrary domains \(D\), however, the KL-eigenpairs have to be computed numerically. The eigenproblem (2.6) can be formulated in variational form as:

\[
\int_D \int_D V_a(x, x') \varphi(x') v(x') \, dx' \, dx = \lambda \int_D \varphi(x) v(x) \, dx.
\]  

For the numerical examples following in the next section, we used a Galerkin discretization with piecewise linear finite element shape functions. The obtained generalized matrix eigenvalue problem was then solved by using JDBSYM, a Jacobi Davidson method optimized for large symmetric eigenproblems, see [26, 25] for details.

### 8.1.3 Localization of quasi-best-\(N\)-term gPC coefficients

The optimal index sets \(\Lambda_{\gamma}(l_1)\) of ‘active’ gPC coefficients are in general not computationally available. However, in the following we will present a heuristic strategy to localize index sets \(\tilde{\Lambda}_{\gamma}(l_1)\) which, in numerical experiments, turn out to be quasi-optimal. Their identification is based on the observation, proved in Section 7.3, that the decay of the KL coefficients of the input random field determines the decay of the coefficients \(u_\nu\) in the expansion (7.8) of the random solution which are to be approximately computed by the stochastic Galerkin algorithm. Precisely, by Lemma 7.3.4 and 7.3.7, respectively, we have that

\[
\|u_\nu\|_{\mathcal{W}(D)} \leq C \eta^{-\nu} := \prod_{m \geq 1} \eta_m^{-\nu_m}
\]
where $\eta_m^{-1} \lesssim \|\psi_m\|_{L^\infty} m^{1+\delta}$ and, by Corollary 2.2.8, $\eta_m^{-1} \lesssim m^{-s+(1+\delta)}$, provided the covariance $V_a$ is in $H^{t,t}(D \times D)$ with $t > 3d$ and $s$ given by (2.26). Hence, the heuristic index sets $\tilde{\Lambda}_\gamma(l_1)$ consist simply of the $\lceil 2^\gamma l_1 \rceil$ largest upper bounds $\eta^{-\nu}$, i.e.

$$\tilde{\Lambda}_\gamma(l_1) := \arg\max_{\Lambda \subset \mathbb{N}^M} \left( \sum_{\nu \in \Lambda} \eta^{-\nu} \right) \subset \mathbb{N}_0^N, \quad l_1 = 0, 1, 2, \ldots \quad (8.2)$$

It has been shown in [8] that the computation of the index sets $\tilde{\Lambda}_\gamma(L)$ can be done linear in time and memory requirement, w.r.t. their size.

### 8.1.4 Postprocessing

Since the goal of our computations is the approximation of the mean value and possibly higher order moments of the solution $\hat{u}_M$, we have to compute expected values of the form $\mathbb{E}[\hat{u}_M], \mathbb{E}[\hat{u}_M^2]$ etc. out of the PC expansion (7.8) or the set of collocated solutions $u(y_j, x)$.

**sCFEM postprocessing**

For the stochastic collocation method, we obtain by means of (6.8)

$$\mathbb{E}\left[I_{l_1}^{(1)}u\right] = \sum_{0 \leq |k| \leq l_1} (-1)^{l_1-|k|} \binom{M-1}{l_1-|k|} \mathbb{E}\left[I_{k_1}^{(1)} \otimes \cdots \otimes I_{k_M}^{(M)}u\right],$$

$$\mathbb{E}\left[I_{l_1}^{(2)}u^2\right] = \sum_{0 \leq |k| \leq l_1} (-1)^{l_1-|k|} \binom{M-1}{l_1-|k|} \mathbb{E}\left[I_{k_1}^{(1)} \otimes \cdots \otimes I_{k_M}^{(M)}u^2\right] \quad (8.3)$$

e
c
e
e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

e
e

sGFEM postprocessing

In order to extract statistical quantities from the sGFEM solution
\[ \hat{u}_M(y, x) = \sum_{\nu \in \Lambda_x(t_1)} u_\nu(x) P_\nu(y), \]
we use the orthogonality of the Legendre and Hermite polynomials, to obtain for the mean value
\[ \mathbb{E}[\hat{u}_M] = u_0(x), \]
where \( 0 = \{0, 0, \ldots, 0\}. \)

Since the variance can be expressed as \( \text{Var}[\hat{u}_M] = \mathbb{E}[\hat{u}_M^2] - \mathbb{E}[\hat{u}_M]^2 \), we obtain the formula
\[ \text{Var}[\hat{u}_M](x) = \sum_{\nu \in \Lambda_x(t_1) \setminus \{0\}} \zeta_\nu u_\nu(x)^2, \]
where \( \zeta_\nu \) denotes the \( L^2(\Gamma) \)-norm of the polynomials \( P_\nu \), i.e.
\[ \zeta_\nu = \left\{ \prod_{m \geq 1} \frac{1}{2m+1} \right\}^{1/\nu} P_\nu = L_\nu \]
\[ P_\nu = H_\nu \]

### 8.2 Numerical examples

The main components of the sparse stochastic Galerkin and sparse stochastic collocation algorithms are the hierarchic discretizations in space and random parameter domain. In the following we will therefore first provide numerical examples for the semi-discretizations in \( D \) and the random parameter space \( \Gamma \), and then for the sparse stochastic Galerkin and collocation algorithms, where we will in particular compare the results to the respective full tensor algorithm.

#### 8.2.1 Wavelet discretization

To illustrate the convergence and the condition of the computed Galerkin stiffness matrix w.r.t. wavelet discretization, we consider the following, purely deterministic model problem
\[ -\text{div}(e^{x_1} \nabla u(x_1, x_2)) = e^{x_1} \pi \cos \left( \frac{3\pi x_2}{2} \right) \left( \cos(\pi x_1) - \frac{13\pi}{4} \sin(\pi x_1) \right) \]
\[ (8.4) \]
8.2 Numerical examples

on $D = [-1, 1]^2$ with zero Dirichlet boundary conditions. The stiffness matrix was computed by a generic finite element code, assembling the matrix w.r.t. the hat basis functions on the finest mesh $T_{l_2}$, in conjunction with a wavelet transform. The solution to (8.4) is $u(x_1, x_2) = \sin(\pi x_1) \cos \left( \frac{3\pi}{2} x_2 \right)$. The initial mesh $T_0$ consists of 200 congruent triangular elements. Figure 8.1 shows the solution obtained after refining the initial mesh twice, i.e. on $T_2$.

Figure 8.2 shows the convergence rate of the wavelet discretization computed up to level $l_2 = 6$, in agreement with (5.7), as well as a comparison of the number of CG-iterations used on each level by a hat function and wavelet discretization, respectively.

8.2.2 Smolyak interpolation

This subsection is concerned with the Smolyak interpolation operator given by (6.7). Motivated by the zero-dimensional problem, i.e. (4.4) without dependence on the spatial variable $x$ (see also [17, 9]),

$$a(y)u(y) = 1 \quad (8.5)$$

with $a(y) = \gamma_0 + \sum_{m=1}^{M} \gamma_m y_m$ and $\gamma_m \in \mathbb{R}$, we consider the problem of computing the expectation $E[u]$ of the solution to (8.5), i.e. the integral

$$I = \int \frac{\rho(y)dy}{\gamma_0 + \sum_{m=1}^{M} \gamma_m y_m} \quad (8.6)$$
by a Smolyak cubature based on Legendre abscissae. Here, we choose $M = 20$ and for $m \geq 1$ we set $\gamma_m = \sqrt{\lambda_m \| \varphi_m \|_{L^\infty(D)}}$ where $(\lambda_m, \varphi_m)$ are the largest $M$ eigenpairs associated to the Gaussian covariance $V_a(x, x') = e^{-\|x-x'\|^2}$, see (2.6). Furthermore we set $\Gamma_m = [-1, 1]$ and $\rho$ to be the product of the uniform probability densities $\rho_m(y_m) = \frac{1}{2}$ on $\Gamma_m$.

The expected rate of convergence, as stated Lemma 6.2.3 and Remark 6.2.4, is strongly dependent on the smallest region of analyticity $\Sigma(\Gamma_m, r_m)$. Therefore, in our case, raising the constant $\gamma_0$ enlarges the domains of analyticity, due to $a_0 \sim \gamma_0$ and (7.32), and should provide better convergence rates. This can clearly be seen in Figure 8.3 where we considered various choices of $\gamma_0$ and where we observe superalgebraic convergence in all cases. The convergence rates have been estimated between two consecutive collocation levels $l_1, l_1+1$ under the assumption that $\epsilon_{l_1} := |I - E[I_{l_1}^{a}]| \sim (N_{l_1}^\Gamma)^{-\gamma}$, where $N_{l_1}^\Gamma$ denotes the number of collocation points at level $l_1$ and $E[I_{l_1}^{a}]$ is computed as in (8.3). Then, the estimated rate $\tilde{r}$ of the convergence is computed by

$$\tilde{r} := \frac{\log(\epsilon_{l_1+1}/\epsilon_{l_1})}{\log(N_{l_1+1}/N_{l_1})}.$$ 

As a second example for the Smolyak procedure, we consider the case of Gaussian random variables, hence the Smolyak algorithm is based on the zeros of Hermite polynomials. Motivated again by the zero-dimensional problem (8.5) with the lognormal coefficient $a = \exp(\sum_{m=1}^M \gamma_m y_m)$, we compute the integral

$$I = \int_{\Gamma} \frac{\rho(y)dy}{e^{\sum_{m=1}^M \gamma_m y_m}} \quad (8.7)$$
8.2 Numerical examples

where $\Gamma = (-\infty, \infty)^M$ and $\rho(y)$ denotes the product of the probability density functions of the standard Gaussian measure. Since the integrand is an entire function, we expect exponential convergence w.r.t the number of collocation points $N_{l_1}^\Gamma$, which can also clearly be seen in Figure 8.4 for different choices of the dimension $M$ as indicated. Under the assumption that $\epsilon_{l_1} := |I - \mathbb{E}[\mathcal{I}_{l_1}]| \sim \exp \left(-r \frac{2M+1}{\sqrt{N_{l_1}}}\right)$, the rate $r$ has been estimated by $\tilde{r}$ between two consecutive collocation levels as

$$\tilde{r} := -\frac{\log(\epsilon_{l_1+1}/\epsilon_{l_1})}{2M+1/N_{l_1+1} - 2M+1/N_{l_1}}.$$  

8.2.3 Sparse tensor stochastic collocation method

To illustrate the sparse tensor stochastic collocation method, we consider the following example problem

$$\begin{cases}
-\text{div}(a(y,x)\nabla u(y,x)) = f(x) & D, \\
u(y,x)|_{x \in \partial D} = 0,
\end{cases}$$

(8.8)

where $D = [-1,1]^2$ and the $y_m$ are uniformly distributed on $\Gamma_m = [-1,1]$. The diffusion coefficient $a$ is given as a truncated Karhunen-Loève expansion

$$a(y,x) = a_M(y,x) = \mathbb{E}_a(x) + \sum_{m=0}^M \sqrt{\lambda_m} \varphi_m(x) y_m$$

(8.9)
8 Implementation and numerical examples

Figure 8.4: Left: Error $\epsilon_{l_1} = |I - E[I_{l_1 \Sigma}]|$ of the Smolyak cubature computing the integral in (8.7). Right: Estimated exponential convergence rate $r$. Both w.r.t. the number of collocation points $N_{l_1}^T$.

with $E_a(x) = 8 + \sin(\pi(x+y))$ and $(\lambda_m, \varphi_m)$ being the largest $M$ computed eigenpairs of the eigenvalue problem (2.6) defined by the covariance $V_a(x, x') = e^{-\|x-x\|^2}$. Finally, we set the source term $f(x) = 2e^{x+2y}$. Figure 8.5 shows the obtained solution for $M = 20$ and on level $L = 2$ of the refinement.

In Figure 8.6 we compare the number of total degrees of freedom $N_L$ and the $L^2(D)$-convergence of the mean $E[u]$ in the sparse stochastic collocation method (4.43) with the expected respective results obtained in the ‘full method’ (4.42) for various choices of $M$ as indicated. Since the Smolyak procedure exhibits a faster than algebraic convergence, we expect a rate of $-\frac{1}{2}$ of the sparse tensor collocation algorithm, coinciding with the rate of the spatial approximations by wavelets, indicated in Figure 8.6.

8.2.4 Sparse tensor stochastic Galerkin method

Again we consider a problem of the form (8.8) on the unit square $D = [-1,1]^2$ with a diffusion coefficient $a$ given by a Karhunen-Loève expansion as in (8.9). To verify the convergence result provided by Proposition 7.2.1, we choose the kernel of the Wiener process as the underlying covariance i.e. $V_a(x, x') = \min(x, x')^2$, where $x \in [-1, 1]$. The eigenvalue problem (2.6) w.r.t. this kernel has the exact solution

$$\tilde{\lambda}_m = \frac{8}{\pi^2(2m-1)^2}, \quad \varphi_m(x) = \sin \left( \frac{x + 1}{\sqrt{2\tilde{\lambda}_m}} \right).$$
8.2 Numerical examples

Figure 8.5: Mean and variance of the solution $u$ to the sPDE described in Section 8.2.3

Figure 8.6: Numerical experiment as described in Section 8.2.3. Left: Comparison between the total number of DoF $N_L$ in the sparse and full stochastic collocation method w.r.t. the level $L$. Right: relative error of the mean $E[u]$ w.r.t. $N_L$. 
hence the eigenvalues are algebraically decaying with rate 2. We then choose $\lambda_m = \tilde{\lambda}^{5/2}_m$, such that $\sqrt{\lambda_m} \|\varphi_m\|_{L^\infty(D)}$ exhibits an algebraic decay with rate $s = \frac{5}{2}$, which, by Proposition 7.1.2, in turn implies a stochastic rate of $r = 1$. Furthermore, we assume $E_a(x) = x_1 + 5$ and $f \equiv 1$. Using piecewise linear wavelets in space corresponds to a spatial approximation rate $\frac{l}{d} = \frac{1}{2}$, provided the solution is accordingly regular. Hence, by Proposition 7.2.1, we expect a sparse tensor rate of $\beta = \frac{1}{2}$ while as in the full tensor case we expect a rate $\bar{\beta} = \frac{1}{3}$, see Remark 7.2.2. Those rates are exactly retrieved by the numerical experiment as observed in Figure 8.7. There, we plot the relative error of the computed solutions in the $L^2_\rho(\Gamma; H^1_0(D))$- and $L^2_\rho(\Gamma; L^2(D))$-norm, respectively, the expected convergence of the Monte Carlo method and the estimated order of convergence, computed between consecutive levels and plotted w.r.t. the total number of degrees of freedom $N_L$. To estimate the order of convergence we assumed

$$
\|u_M\|_{L^2_\rho(\Gamma; H^1_0(D))}^2 - \|\hat{u}_M\|_{L^2_\rho(\Gamma; H^1_0(D))}^2 \sim \dim(V^T_L \otimes V^D_L)^{-2\beta}
$$

and

$$
\|u_M\|_{L^2_\rho(\Gamma; H^1_0(D))}^2 - \|\bar{u}_M\|_{L^2_\rho(\Gamma; H^1_0(D))}^2 \sim \dim(V^T_L \otimes V^D_L)^{-2\bar{\beta}},
$$

where $u_M$ is the exact solution to (4.5), $\hat{u}_M$ the discrete solution to (4.25) w.r.t the sparse tensor product space $V^T_L \otimes V^D_L$ and $\bar{u}_M$ the discrete solution to (4.22) w.r.t the full tensor product space $V^T_L \otimes V^D_L$. The three unknown quantities in each relation, i.e. the solution $u_M$, the rates $\beta$ and $\bar{\beta}$, respectively and the constant are then fitted using discrete solutions on three consecutive levels.
Figure 8.7: Numerical experiment as described in Section 8.2.4. Left: Convergence of the solutions computed by a sparse tensor Galerkin (STG), full tensor Galerkin (FTG) and Monte Carlo (MC) method. Right: Corresponding estimated orders of convergence. Both plotted versus the total number of degrees of freedom $N_L$ in the full and sparse tensor case, respectively.
References


References


References


Curriculum Vitae

Personal details

Name  
Marcel Bieri

Date of birth  
January 7, 1979

Place of birth  
Basel, Switzerland

Citizenship  
Switzerland

Education

PhD studies in Mathematics at ETH Zürich
Zurich, Switzerland

09/2004  
Diploma in Mathematics at ETH Zürich

10/1999–09/2004  
Studies in Mathematics at ETH Zürich
Zürich, Switzerland

12/1998  
Matur at Gymnasium Muttenz

Secondary school at Gymnasium Muttenz
Muttenz, Switzerland