Numerical Methods for Elliptic Partial Differential Equations with Random Coefficients

Author: Chiara Pizzigoni

Supervisor: dr. S.G. Cox
            prof. dr. R.P. Stevenson

Examination date: March 9, 2017
Abstract

This thesis analyses the stochastic collocation method for approximating the solution of elliptic partial differential equations with random coefficients. The method consists of a finite element approximation in the spatial domain and a collocation at the zeros of suitable tensor product orthogonal polynomials in the probability space and naturally leads to the solution of uncoupled deterministic problems. The computational cost of the method depends on the choice of the collocation points and thus we compare few possible constructions. Although the random fields describing the coefficients of the problem are in general infinite-dimensional, an approximation with certain optimality properties is obtained by truncating the Karhunen-Loève expansion of these random fields. We estimate the convergence rates of the method, depending on the regularity of the random coefficients. In particular we prove exponential convergence in probability space. Numerical examples illustrate the theoretical results.

Title: Numerical Methods for Elliptic Partial Differential Equations with Random Co-efficients
Author: Chiara Pizzigoni, chiarapizzigoni92@gmail.com, 10978801
Supervisor: dr. S.G. Cox, prof. dr. R.P. Stevenson
Examination date: March 9, 2017

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 105-107, 1098 XG Amsterdam
http://kdvi.uva.nl
# Contents

Acknowledgements .............................................. v
Introduction .................................................. vii

1 Preliminaries ............................................. 1
  1.1 The Karhunen-Loève Expansion ......................... 1
    1.1.1 Existence of the Karhunen-Loève Expansion ......... 1
    1.1.2 Truncation of the Karhunen-Loève Expansion ....... 7
  1.2 The Karhunen-Loève Expansion of a Random Field .......... 10
    1.2.1 More on Convergence of the Karhunen-Loève Expansion . 12

2 Problem Setting ........................................... 15
  2.1 Strong and Weak Formulations ......................... 15
  2.2 Finite-Dimensional Stochastic Space .................... 17

3 The Stochastic Collocation Method ....................... 19
  3.1 Full Tensor Grid of Collocation Points ................. 19
  3.2 Sparse Tensor Grid of Collocation Points ............. 21
    3.2.1 Anisotropic Version ................................ 24

4 Convergence of the Method ................................ 27
  4.1 Regularity Results ................................... 27
  4.2 Convergence Analysis .................................. 35
    4.2.1 Error of the Truncation ......................... 35
    4.2.2 Finite Element Error ........................... 36
    4.2.3 Collocation Error - Full Tensor Method ........... 37
    4.2.4 Convergence Result ............................ 47
    4.2.5 Collocation Error - Sparse Tensor Method ......... 48
  4.3 Convergence of Moments ................................ 51
  4.4 Anisotropic Sparse Method - Selection of Weights $\alpha$ ... 52

5 Numerical Implementation ................................ 53
  5.1 Example .............................................. 53
  5.2 Conclusions .......................................... 60

Summary ...................................................... 61

Bibliography .................................................. 63
Acknowledgements

I am more than grateful to Sonja and Rob. Thank you for supporting me and not only supervising. In Italian the word support is translated with supporto and if you change only one letter getting sopporto, the meaning will become the slightly more negative tolerate. So I really have to thank you for both supporting and stand me. Thank you for the time, the energies and the attention that both of you have dedicated to me.

When the mathematics is not the biggest problem, I can always rely on them. Thank you Sara for enlightening when I can see nothing but darkness. Thank you Marco for being always next to me even if you are far away. Thanks to my soul mates, Francesca, Linda and Alessia, loyal mates and beautiful souls.

Silvia, thank you for being simply the best person standing by my side and for giving your all to me.

I personally think that behind great men and women there are always great parents. So thank you mum and dad for being great even if I am not. Grazie!
Elliptic second order partial differential equations are well suited to describe a wide variety of phenomena which present a static behaviour, e.g. the stationary solution to a diffusion problem. An entire branch of research is dedicated to the theoretical analysis and numerical implementation of methods which allow to approximate the exact solution of these equations. On the other hand one should always keep in mind that in practice these models are approximations of the physical system and they do not describe the given problem exactly. Therefore we aim to study equations which account as much as possible for the uncertainties arising naturally from the mathematical observation of the real world. Sources of such uncertainties can be for example errors in the measurements, the intrinsic nature of the system, partially known data sets or an overall knowledge extrapolated from only a few spatial locations.

The mathematical model describing a phenomenon can be thought as an input-output machine which takes some functions and returns the solution of the model. Hence we expect the inaccuracies in the inputs to easily propagate to the output.

In order to include in our mathematical examination all these moderately predictable factors, we model them as noises following some probability distribution. Thus we turn our attention to partial differential equations where the coefficients are random fields depending on these uncertain parameters. Namely, instead of focusing on elliptic equations, defined on a spatial domain $D \subset \mathbb{R}^d$, of the form

$$-\nabla \cdot (a(x) \nabla u(x)) = f(x), \quad x \in D,$$

we consider the following model

$$-\nabla \cdot (a(\omega, x) \nabla u(\omega, x)) = f(\omega, x), \quad (\omega, x) \in \Omega \times D \quad (0.0.1)$$

where $\Omega$ is the set of possible outcomes. Again the main difference relies on the fact that the functions of the second problem have a stochastic representation emphasizing the inexactness of the coefficients of which we suppose to know the law.

In our discussion we consider a particular parametrization of the random coefficients given by the Karhunen-Loève expansion, a linear combination of infinitely many uncorrelated random variables. This choice is motivated because it guarantees some optimal results as it has been proved in \[14\]. We underline the fact that in general this decomposition is infinite-dimensional. The computational necessity to deal with finite-dimensional objects leads to the need of truncating the Karhunen-Loève expansion and the quantification of the consequent error.

To approximate numerically the solution of problem (0.0.1) appended with some suitable boundary conditions, we will use the stochastic collocation method which employs
standard deterministic techniques in the spatial domain and tensor product polynomial approximation in the random domain. This method is extensively analysed in [11]. Unfortunately tensor product spaces suffer from the so-called curse of dimensionality as the dimension of the polynomial space grows exponentially fast in the number of terms that we keep in the Karhunen-Loève truncation. In case where this number becomes quite large we may switch to sparse tensor product spaces which highly reduce the computational complexity of the method while preserving a good effectiveness (see [12]).

Concretely the procedure consists in approximating, using the Galerkin finite element scheme, the solutions of problems which are deterministic in the domain $D$ once we evaluate the probabilistic variables at several collocation points. The final approximation is then recovered by interpolating the semi-discrete finite element approximations. This approach differs from the widely used Monte Carlo method in the selection of the evaluation points. The latter adopts a random pick subject to a given distribution. By applying the stochastic collocation method instead, the points are chosen as the roots of (possibly sparse) tensor product polynomials orthogonal with respect to the joint probability density of the random variables appearing in the Karhunen-Loève truncation. This preference benefits from the nice properties of the zeros of orthogonal polynomials (see [15], Chapter III). While keeping the advantage of solving uncoupled deterministic problems as in the Monte Carlo approach, the stochastic collocation achieves a faster convergence rate as we will prove in Chapter 4.

The main goals of this work are to provide a rigorous investigation on the existence of the Karhunen-Loève parametrization and how its truncation influences our analysis, to describe the numerical techniques to solve a stochastic model problem and to estimate in detail the errors arising from the subsequent approximations that we introduce.

The outline of the thesis is as follows: Chapter 1 focuses on the existence of the Karhunen-Loève expansion of an element in a tensor product Hilbert space and on the estimation of the error, measured in some suitable norm, which arises as a consequence of truncating the expansion. Afterwards this abstract setting is tuned to our specific situation, analysing the results in the special case of random fields.

Chapter 2 describes the infinite-dimensional boundary value problem whose solution we want to approximate. In particular we show that such a solution indeed exists and in order to give an approximation, we turn to the corresponding finite-dimensional problem where all the random fields are represented by truncations of the Karhunen-Loève decomposition.

In Chapter 3 we present the stochastic collocation method, both in its full and sparse versions, and briefly discuss the anisotropic (direction dependent) variant of the last one.

The aim of Chapter 4 is to provide a rigorous error estimate of the entire method by splitting the error into three parts: the truncation error, the approximation error in the spatial domain and the approximation error in the stochastic space. In particular the latter is analysed both for the full and sparse version of the method comparing the convergence rate of the two. This is done under some mild regularity assumptions on the coefficients entering the problem which we need to impose in order to ensure cor-
responding regularity of the solution in the random domain. The theoretical result con-
cerning the collocation error, which is obtained conducting a stepwise one-dimensional
analysis, guarantees exponential convergence in the random space for both variants of
the stochastic collocation method.

Chapter 5 is devoted to some computational examples including a numerical com-
parison with the Monte Carlo method.
1 Preliminaries

1.1 The Karhunen-Loève Expansion

Uncertainties in a physical model are very often modelled as random fields. The aim of this section is to prove that any infinite-dimensional random field can be represented by the so-called Karhunen-Loève expansion which is roughly speaking an infinite linear combination of uncorrelated random variables. Moreover the best (in terms of mean square error) finite-dimensional approximation of the random field is obtained by truncating this expansion. We will give precise bounds for this error. Following [14], we develop our construction in the general framework of Hilbert spaces. Throughout this chapter all Hilbert spaces are real and separable unless stated otherwise.

1.1.1 Existence of the Karhunen-Loève Expansion

Let \((H_1, \langle \cdot, \cdot \rangle_{H_1}), (H_2, \langle \cdot, \cdot \rangle_{H_2})\) and \((S, \langle \cdot, \cdot \rangle_S)\) be separable Hilbert spaces over \(\mathbb{R}\). For \(i \in \{1, 2\}\), let \((e^{(i)}_n)_{n \in I_i}\) and \((s_m)_{m \in J}\) be orthonormal bases of \(H_i\) and \(S\), respectively, with \(I_i\) and \(J\) countable sets. For \(x \in H_i\) and \(y \in S\) define the bilinear form \(x \otimes y : H_i \times S \to \mathbb{R}\)

\[
[x \otimes y](a, b) := \langle x, a \rangle_{H_i} \langle y, b \rangle_S, \quad (a, b) \in H_i \times S.
\]

Let \(E\) be the set of all finite linear combinations of such bilinear forms. We define the inner product \(\langle \cdot, \cdot \rangle_{H_i \otimes S}\) on \(E\) as follows

\[
\langle x_1 \otimes y_1, x_2 \otimes y_2 \rangle_{H_i \otimes S} := \langle x_1, x_2 \rangle_{H_i} \langle y_1, y_2 \rangle_S
\]

where \(x_1, x_2 \in H_i\) and \(y_1, y_2 \in S\).

Definition 1.1.1. The tensor product of \(H_i\) and \(S\) denoted by \(H_i \otimes S\) is defined as the completion of \(E\) under the inner product \(\langle \cdot, \cdot \rangle_{H_i \otimes S}\).

It can be shown that \((e^{(i)}_n \otimes s_m)_{(n,m) \in I_i \times J}\) is an orthonormal basis for the space \(H_i \otimes S\) [1, Section II.4]. Thus we can represent any element \(f \in H_i \otimes S\) as

\[
f = \sum_{(n,m) \in I_i \times J} c_{n,m} e^{(i)}_n \otimes s_m
\]

where \(c_{n,m} = \langle f, e^{(i)}_n \otimes s_m \rangle_{H_i \otimes S}\) and \(\sum_{(n,m) \in I_i \times J} c^2_{n,m} = \|f\|^2_{H_i \otimes S}\).

Define \(f_m := \sum_{n \in I_i} c_{n,m} e^{(i)}_n \in H_i\). It is useful for later use to observe that \(\|f_m\|_{H_i}^2 = \sum_{n \in I_i} c^2_{n,m}\). We get the following expansion for any element \(f \in H_i \otimes S\):

\[
f = \sum_{m \in J} f_m \otimes s_m.
\]
We can prove the following

**Proposition 1.1.2.** Let \( f \in H_1 \otimes S \) and \( g \in H_2 \otimes S \). The map \( C_{.,.} : (H_1 \otimes S) \times (H_2 \otimes S) \rightarrow H_1 \otimes H_2 \) defined as

\[
C_{fg} = \sum_{m \in J} f_m \otimes g_m
\]

is bilinear, well-defined and independent on the choice of the orthonormal basis in \( S \).

**Proof.** Bilinearity comes directly from the distributive property of tensor product. We prove that the map is well-defined:

\[
\|C_{fg}\|_{H_1 \otimes H_2} \leq \left( \sum_{m \in J} \|f_m\|_{H_1}^2 \right)^{\frac{1}{2}} \left( \sum_{m \in J} \|g_m\|_{H_2}^2 \right)^{\frac{1}{2}} = \|f\|_{H_1 \otimes S} \|g\|_{H_2 \otimes S}
\]

where we have used the Cauchy-Schwarz inequality.

Finally we show that the map is independent on the choice of the orthonormal basis in \( S \). Let \( \{s_m\}_{m \in J} \) be an orthonormal basis of \( S \) and \( \{s'_m\}_{m \in J} \) be a second one. Consequently for each \( m \in J \) there exist \( (\gamma_{n,m})_{n \in J} \subset \mathbb{R} \) such that \( s'_m = \sum_{n \in J} \gamma_{n,m} s_n \). By the orthonormality condition it follows that

\[
\sum_{n \in J} \gamma_{n,m} \gamma_{n,k} = \delta_{mk}.
\]

Indeed

\[
\delta_{mk} = (s'_m, s'_k)_S = \sum_{n \in J} \gamma_{n,m} s_n \sum_{l \in J} \gamma_{l,k} s_l)_S = \sum_{(n,l) \in J \times J} \gamma_{n,m} \gamma_{l,k} (s_n, s_l)_S = \sum_{n \in J} \gamma_{n,m} \gamma_{n,k}.
\]

Let \( \{e_n^{(i)}\}_{n \in I_i} \) be an orthonormal basis of \( H_i \). As \( f \in H_1 \otimes S \) there exist \( \{\alpha_{n,m}\}_{(n,m) \in I_1 \times J} \subset \mathbb{R} \) such that

\[
f = \sum_{(n,m) \in I_1 \times J} \alpha_{n,m} e_n^{(1)} \otimes s_m.
\]

Similarly as \( g \in H_2 \otimes S \) there exist \( \{\beta_{n,m}\}_{(n,m) \in I_2 \times J} \subset \mathbb{R} \) such that

\[
g = \sum_{(n,m) \in I_2 \times J} \beta_{n,m} e_n^{(2)} \otimes s_m.
\]

Now we expand \( f \) and \( g \) with respect to the basis \( \{s'_m\}_{m \in J} \):

\[
f = \sum_{(n,k) \in I_1 \times J} \alpha'_{n,k} e_n^{(1)} \otimes s'_k = \sum_{k \in J} \sum_{(n,m) \in I_1 \times J} (\alpha'_{n,k} \gamma_{n,k}) e_n^{(1)} \otimes s_m.
\]
\[ g = \sum_{(n,k) \in I_2 \times J} \beta'_{n,k} e_n^{(2)} \otimes s_k' = \sum_{k \in J} \sum_{(n,m) \in I_2 \times J} (\beta'_{n,k} \gamma_{m,k}) e_n^{(2)} \otimes s_m. \]

By uniqueness of the expansion it holds

\[ \alpha_{n,m} = \sum_{k \in J} \alpha'_{n,k} \gamma_{m,k}, \quad \beta_{n,m} = \sum_{k \in J} \beta'_{n,k} \gamma_{m,k}. \]

Therefore we have

\[
C_{f,g}^{(s_m)} = \sum_{m \in J} \left( \sum_{n \in I_1} \alpha_{n,m} e_n^{(1)} \right) \otimes \left( \sum_{l \in I_2} \beta_{l,m} e_l^{(2)} \right)
= \sum_{(n,l,m) \in I_1 \times I_2 \times J} (\alpha_{n,m} \beta_{l,m}) e_n^{(1)} \otimes e_l^{(2)}
= \sum_{(n,l) \in I_1 \times I_2} \left( \sum_{k \in J} \alpha'_{n,k} \beta'_{l,i} \gamma_{m,k} \gamma_{m,i} \right) e_n^{(1)} \otimes e_l^{(2)}
= \sum_{(n,l) \in I_1 \times I_2} \left( \sum_{(k,i) \in J \times J} \alpha'_{n,k} \beta'_{l,i} \delta_{ki} \right) e_n^{(1)} \otimes e_l^{(2)}
= \sum_{k \in J} \left( \sum_{n \in I_1} \alpha'_{n,k} e_n^{(1)} \right) \otimes \left( \sum_{l \in I_2} \beta'_{l,k} e_l^{(2)} \right) = C_{f,f}^{(s_m)}. \]

As a consequence of the previous result we are allowed to introduce the following

**Definition 1.1.3.** $C_{f,g}$ defined in Proposition 1.1.2 is called the correlation of $f$ and $g$.

Now we investigate the possible existence of an operator which can be associated to the correlation

\[ C_f := C_{f,f} \]

for $f \in H \otimes S$. Before doing this we recall some concepts and results of functional analysis. Let $C : H \to H$ be an operator on the real Hilbert space $H$. We can associate to
The norm
\[ \|C\| := \sup_{v \in H, \|v\|_H = 1} \|Cv\|_H. \]
We say that a linear and bounded operator \( C \) on \( H \) is compact if there exists a sequence \((C_n)\) of finite rank operators such that
\[ \|C - C_n\| \to 0. \]

The Spectral Theorem for compact and symmetric operators on a separable Hilbert space states (see [2], Theorem 5.1)

**Theorem 1.1.4.** Let \( H \) be a separable real Hilbert space. Let \( C \) be a symmetric and compact operator on \( H \). Then for any \( v \in H \)
\[ Cv = \sum_{m \in J} \lambda_m \langle v, e_m \rangle_H e_m \]
where \((e_m)_m\) form an orthonormal basis for \( H \), \((\lambda_m)_m \subset \mathbb{R}\) are the eigenvalues corresponding to the eigenvectors \( e_m \) of \( C \) and \( \lambda_m \downarrow 0 \).

The notation \( \lambda_m \downarrow 0 \) denotes a non-increasing sequence converging to 0.

Define, for any orthonormal basis \((e_m)_m \in J\), the trace of the non-negative definite symmetric operator \( C \) as
\[ \text{Tr}(C) := \sum_{m \in J} \langle Ce_m, e_m \rangle_H. \]
Indeed it can be proved that this definition is independent on the choice of the basis in \( H \), see for example [1]. We say that a non-negative definite symmetric compact operator is trace class if \( \text{Tr}(C) < \infty \). Equivalently by using the characterization of the Spectral Theorem, a non-negative definite symmetric compact operator is trace class if \( \sum_m \lambda_m < \infty \).

We can now proceed to prove the following

**Theorem 1.1.5.** Let \((H, \langle \cdot, \cdot \rangle_H)\) and \((S, \langle \cdot, \cdot \rangle_S)\) be separable Hilbert spaces of the same dimension and let \((e_m)_{m \in J}, (s_m)_{m \in J}\) be orthonormal bases of \( H \) and \( S \), respectively. The map \( \Phi : \{C_f \in H \otimes H : f \in H \otimes S\} \to \{C : C \text{ non-negative definite trace class operator}\} \) given by
\[ \Phi(C_f)(v) = \Phi \left( \sum_{m \in J} f_m \otimes f_m \right)(v) := \sum_{m \in J} \langle f_m, v \rangle_H f_m \in H, \quad v \in H \quad (1.1.1) \]
is a one-to-one correspondence.

**Proof.** For \( f \in H \otimes S \) we denote \( \Phi(C_f) \) by \( C_f \). First we prove that indeed \( C_f \) has the required properties. By definition we can immediately conclude that \( C_f \) is compact as it is a norm limit of finite-rank operators. Indeed if we define \( C_n := \sum_{m \leq n} \langle f_m, \cdot \rangle_H f_m \) then
\[ \|C_f - C_n\| \leq \sum_{m > n} \|f_m\|^2 \to 0, \quad n \to \infty. \]
We show that $C_f$ is non-negative definite. Let $v \in H, v \neq 0$. Then
\[
(C_f v, v)_H = \sum_{m \in J} (f_m, v)_H (f_m, v)_H = \sum_{m \in J} (f_m, v)_H^2 \geq 0.
\]

Now we check that $C_f$ is a trace class operator. Namely plugging the expression for $C_f$ into the definition of trace,
\[
\text{Tr}(C_f) = \sum_{m \in J} (C_f e_m, e_m)_H = \sum_{m \in J} \sum_{n \in J} (f_n, e_m)_H (f_m, e_n)_H = \sum_{n \in J} \sum_{m \in J} (f_n, e_m)_H^2 (e_m, e_m)_H
\]
\[
= \sum_{n \in J} (\sum_{m \in J} (f_n, e_m)_H e_m, \sum_{m \in J} (f_n, e_m)_H e_m)_H = \sum_{n \in J} \|f_n\|_H^2 = \|f\|_{H \otimes S}^2 < \infty.
\]

Therefore we can conclude that $C_f$ is a non-negative definite trace class operator.

Now it remains to show that the map $\Phi$ is a one-to-one correspondence. Observe that the following chain of equalities holds true for $v, w \in H$:
\[
(C_f v, w)_H = \sum_{m \in J} (f_m, v)_H (f_m, w)_H = \sum_{m \in J} (f_m \otimes f_m, v \otimes w)_{H \otimes H}
\]
\[
= (\sum_{m \in J} f_m \otimes f_m, v \otimes w)_{H \otimes H} = (C_f, v \otimes w)_{H \otimes H}. \tag{1.1.2}
\]

We are going to check that our map is injective, i.e. $C_f = C_g$ for $f, g \in H \otimes S$ implies $C_f = C_g$. Assuming that $C_f = C_g$ we have by (1.1.2) that for all $v, w \in H$
\[
(C_f, v \otimes w)_{H \otimes H} = (C_g, v \otimes w)_{H \otimes H}.
\]

By linearity we have also for all $n \in \mathbb{N}$ and for all $v_i, w_i \in H$
\[
(C_f, \sum_{i=1}^n v_i \otimes w_i)_{H \otimes H} = (C_g, \sum_{i=1}^n v_i \otimes w_i)_{H \otimes H}
\]
or equivalently
\[
(C_f - C_g, \sum_{i=1}^n v_i \otimes w_i)_{H \otimes H} = 0.
\]

The claim is proved after observing that the set $\{\sum_{i=1}^n v_i \otimes w_i : v_i, w_i \in H, n \in \mathbb{N}\}$ is a dense subset in $H \otimes H$ and therefore $C_f - C_g = 0$.

It remains to show that the map is also surjective. Let $C$ be a non-negative definite trace class operator. We want to show that there exist $f \in H \otimes S$ such that $C_f = C$. Let $(\phi_m)_{m \in J}$ be the sequence of eigenvectors of $C$ forming an orthonormal basis for $H$ and $(\lambda_m)_{m \in J} \subset \mathbb{R}_+$ be the eigenvalues of $C$, i.e.
\[
C \phi_m = \lambda_m \phi_m. \tag{1.1.3}
\]

Moreover $\sum_m \lambda_m < \infty$ as $C$ is trace class. As a consequence we get convergence of the following series
\[
f = \sum_{m \in J} \sqrt{\lambda_m} \phi_m \otimes s_m.
\]
For this element we have $f_m = \sqrt{\lambda_m} \phi_m$. Hence

$$C_f = \sum_{m \in J} \lambda_m \phi_m \otimes \phi_m.$$ 

Thus for all $v \in H$ it holds that

$$C_f v = \sum_{m \in J} \lambda_m \langle \phi_m, v \rangle_H \phi_m.$$ 

From this we can state that the spectrum of $C_f$ equals (1.1.3) and therefore $C_f = C$. 

**Corollary 1.1.6.** Let $(H, \langle \cdot, \cdot \rangle_H)$ be a separable Hilbert space and let $C \in H \otimes H$ be a correlation. Let $C$ be defined by (1.1.1) with eigenpairs $(\lambda_m, \phi_m)_{m \in J}$. Then $C$ can be represented as

$$C = \sum_{m \in J} \lambda_m \phi_m \otimes \phi_m. \quad (1.1.4)$$

The following theorem is crucial for our purpose.

**Theorem 1.1.7.** Let $(H, \langle \cdot, \cdot \rangle_H)$ and $(S, \langle \cdot, \cdot \rangle_S)$ be separable Hilbert spaces of the same dimension. Let $(\phi_m)_{m \in J}$ be an orthonormal basis for $H$. Let $C \in H \otimes H$ be a correlation represented as in (1.1.4). Let $f \in H \otimes S$. Then $C_f = C$ if and only if there exists an orthonormal family $(Y_m)_{m \in J} \subset S$ such that

$$f = \sum_{m \in J} \sqrt{\lambda_m} \phi_m \otimes Y_m. \quad (1.1.5)$$

**Proof.** Assume that there exists an orthonormal family $(Y_m)_{m \in J} \subset S$ such that $f = \sum_m \sqrt{\lambda_m} \phi_m \otimes Y_m$. Then by Proposition 1.1.2 we can conclude the statement, possibly after having completed the family $(Y_m)_{m \in J}$ to a basis for $S$.

On the other way around, assume that $C_f = C$. We have already observed at the beginning of the section that we can represent the element $f \in H \otimes S$ as

$$f = \sum_{m \in J} \phi_m \otimes X_m \quad (1.1.6)$$

where $(X_m)_{m \in J} \subset S$. Moreover we have the expansion

$$X_m = \sum_{n \in J} \langle X_m, s_n \rangle_S s_n$$

for $(s_n)_{n \in J}$ being an orthonormal basis of $S$. Thus we get

$$f = \sum_{m \in J} \phi_m \otimes X_m = \sum_{m \in J} \sum_{n \in J} \langle X_m, s_n \rangle_S \phi_m \otimes s_n.$$
If we call \( f_n := \sum_{m \in J} \langle X_m, s_n \rangle S \phi_m \) we have

\[
C_f = \sum_n f_n \otimes f_n = \sum_{n \in J} \left( \sum_{m \in J} \langle X_m, s_n \rangle S \right) \left( \sum_{m' \in J} \langle X_{m'}, s_n \rangle S \right) \phi_m \otimes \phi_{m'}
\]

\[
= \sum_{(m, m') \in J \times J} \left( \sum_{n \in J} \langle X_m, s_n \rangle S \langle X_{m'}, s_n \rangle S \right) \phi_m \otimes \phi_{m'}
\]

\[
= \sum_{(m, m') \in J \times J} \langle X_m, X_{m'} \rangle S \phi_m \otimes \phi_{m'}.
\]

If we compared the previous result with (1.1.4) we conclude that necessarily

\[
\langle X_m, X_{m'} \rangle S = \lambda_m \delta_{mm'}.
\]

Therefore \((X_m)_m\) is an orthogonal family with \( \|X_m\|_S^2 = \lambda_m \). Hence from (1.1.6) we get the statement with \( Y_m = \frac{X_m}{\sqrt{\lambda_m}} \).

**Definition 1.1.8.** The representation (1.1.5) of \( f \) given the spectrum of \( C_f \) is called the Karhunen-Loève expansion of \( f \).

### 1.1.2 Truncation of the Karhunen-Loève Expansion

For any element \( f \in H \otimes S \) with a given correlation we have established the existence of an expansion taking the form

\[
f = \sum_{m \in J} \sqrt{\lambda_m} \phi_m \otimes Y_m
\]

where \((\phi_m)_{m \in J}\) and \((Y_m)_{m \in J}\) are orthonormal systems and \( \lambda_m \downarrow 0 \). Now we are interested in investigating the properties of such an expansion. In order to do that we introduce the following standard notation: for \( U \) a closed subspace of \( H \), \( P_U \) will indicate the orthogonal projection of \( H \) onto \( U \).

**Theorem 1.1.9.** If \( f \in H \otimes S \) has the Karhunen-Loève expansion (1.1.5), then for any \( N \in \mathbb{N} \) it holds

\[
\inf_{U \subset H} \dim_U \sup_{U \subset H} \|f - P_U \otimes S f\|_{H \otimes S}^2 \geq \sum_{m \geq N+1} \lambda_m, \quad (1.1.7)
\]

with equality in the case \( U = \text{span}\{\phi_1, ..., \phi_N\} \).
Proof. If $U = \text{span}\{\phi_1, \ldots, \phi_N\}$ then $P_{U \otimes S}f = \sum_{m=1}^{N} \sqrt{\lambda_m} \phi_m \otimes Y_m$ which is exactly the $N^{\text{th}}$ truncation of the Karhunen-Loève expansion. Thus

$$
\| f - P_{U \otimes S}f \|_{H \otimes S}^2 = \| \sum_{m \geq N+1} \sqrt{\lambda_m} \phi_m \otimes Y_m \|_{H \otimes S}^2 = \sum_{m \geq N+1} \lambda_m \langle \phi_m \otimes Y_m \rangle = \sum_{m \geq N+1} \lambda_m
$$

and so we have equality in (1.1.7).

We prove the first statement by induction. If $N = 0$ then (1.1.7) holds. Firstly observe that we are allowed to take $(Y_m)_{m \in J}$ to be an orthonormal basis of $S$ with $J$ finite or countable infinite. Let $U \subset H$ be closed and such that $\dim U = N$. Consider $g \in U \otimes S$. Then we can represent $g$ as

$$
g = \sum_{m \in J} u'_m \otimes Y_m
$$

where $(u'_m)_m \subset U$. It holds that

$$
g = \sum_{m \in J} \sqrt{\lambda_m} u_m \otimes Y_m
$$

with $u_m := \frac{u'_m}{\sqrt{\lambda_m}} \subset U$. Using this we get

$$
\| f - g \|_{U \otimes H}^2 = \langle f - g, f - g \rangle_{U \otimes H} = \langle \sum_{m \in J} \sqrt{\lambda_m} (\phi_m - u_m) \otimes Y_m, \sum_{n \in J} \sqrt{\lambda_n} (\phi_n - u_n) \otimes Y_n \rangle_{U \otimes H} = \sum_{m \in J} \sum_{n \in J} \sqrt{\lambda_m} \sqrt{\lambda_n} \langle \phi_m - u_m \otimes Y_m, (\phi_n - u_n) \otimes Y_n \rangle_{U \otimes H} = \sum_{m \in J} \sum_{n \in J} \sqrt{\lambda_m} \sqrt{\lambda_n} \langle \phi_m - u_m \otimes Y_m, \phi_n - u_n \rangle_Y H \langle Y_m, Y_n \rangle_S = \sum_{m \in J} \lambda_m \| \phi_m - u_m \|_{H}^2.
$$

(1.1.8)

Consider $\text{span}\{u_1, \ldots, u_{N-1}\} \subset U$. If $\dim(\text{span}\{u_1, \ldots, u_{N-1}\}) = N - 1$ then

$$
W := \text{span}\{u_1, \ldots, u_{N-1}\}.
$$

Otherwise if $\dim(\text{span}\{u_1, \ldots, u_{N-1}\}) < N - 1$ take

$$
W := \text{span}\{u_1, \ldots, u_{N-1}, \phi_N, \ldots, \phi_{N+k}\}.
$$
for some $k \in \mathbb{N}$ such that $\dim W = N - 1$. We use the notation $W^\perp$ for the space such that $U = W \oplus (W^\perp)$. Observe that $\dim(W^\perp) = 1$ and therefore $W^\perp = \text{span}\{e\}$ for some $e \in U$ with $\|e\|_H = 1$. We have

$$
\|f - g\|_{U \otimes H}^2 = \sum_{m \leq N - 1} \lambda_m \|\phi_m - u_m\|_H^2 + \sum_{m \geq N} \lambda_m \|\phi_m - u_m\|_H^2 \\
\quad = \sum_{m \leq N - 1} \lambda_m \|\phi_m - P_W u_m\|_H^2 + \sum_{m \geq N} \lambda_m \|\phi_m - u_m\|_H^2 + \\
\quad + \sum_{m \geq N} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \sum_{m \geq N} \lambda_m \|\phi_m - P_W u_m\|_H^2 \\
\quad = \sum_{m \in J} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \sum_{m \geq N} \lambda_m (\|\phi_m - P_W u_m\|_H^2 - \|\phi_m - u_m\|_H^2) \\
\quad \overset{(*)}{\geq} \sum_{m \in J} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \sum_{m \geq N} \lambda_m \|P_{W^\perp} \phi_m\|_H^2 \\
\quad = \sum_{m \in J} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \sum_{m \geq N} \lambda_m (\|e, \phi_m\|_H^2) \\
\quad \geq \sum_{m \in J} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \sup_{m \geq N} \lambda_m \sum_{m \geq N} (e, \phi_m)_H^2 \\
\quad \geq \sum_{m \in J} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \lambda_N \|e\|_H^2 \\
\quad = \sum_{m \in J} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \lambda_N
$$

where in the last inequality we have used Bessel’s inequality and the assumption that the sequence of eigenvalues is non-increasing. The inequality marked with $(*)$ holds true because

$$
\|\phi_m - P_W u_m\|_H^2 = \|P_W \phi_m - P_W u_m + P_{W^\perp} \phi_m\|_H^2 \\
\quad = \|P_W \phi_m - u_m\|_H^2 + \|P_{W^\perp} \phi_m\|_H^2 \\
\quad = \|P_W \phi_m - u_m\|_H^2 + \|P_{W^\perp} \phi_m\|_H^2 \\
\quad \leq \|\phi_m - u_m\|_H^2 + \|P_{W^\perp} \phi_m\|_H^2.
$$

By (1.1.2) for $w = \sum_{m \in J} \sqrt{\lambda_m} P_W u_m \otimes Y_m \in W \otimes S$ we have

$$
\sum_{m \in J} \lambda_m \|\phi_m - P_W u_m\|_H^2 - \lambda_N = \|f - w\|_{W \otimes S}^2 - \lambda_N \geq \inf_{h \in W \otimes S} \|f - h\|_{W \otimes S}^2 - \lambda_N.
$$

Thus we have shown

$$
\|f - g\|_{U \otimes H}^2 \geq \inf_{W \subseteq H} \inf_{h \in W \otimes S} \|f - h\|_{W \otimes S}^2 - \lambda_N.
$$
Applying the inductive hypothesis we finally get

\[
\inf_{U \subset \text{closed}, \dim U = N} \inf_{g \in U \otimes S} \|f - g\|_H^2 \geq \inf_{W \subset \text{closed}, \dim W = N - 1} \inf_{h \in W \otimes S} \|f - h\|_S^2 - \lambda_N
\]

\[
\geq \sum_{m \geq N} \lambda_m - \lambda_N
\]

\[
= \sum_{m \geq N + 1} \lambda_m.
\]

\[\square\]

1.2 The Karhunen-Loève Expansion of a Random Field

The main goal of this section is to specialize the results we have previously obtained to the particular case of infinite-dimensional random fields. In the notation of the previous section, let \(H = L^2(D)\) and \(S = L^2_P(\Omega)\). Let \(a \in L^2_P(\Omega) \otimes L^2(D)\) (i.e. \(a\) is a second order random field) with mean

\[
E_a(x) := \int \Omega a(\omega, x) dP(\omega)
\]

and covariance function

\[
V_a(x, x') := \int \Omega a(\omega, x)a(\omega, x') dP(\omega) - E_a(x)E_a(x')
\]

\[
= \int \Omega [a(\omega, x) - E_a(x)][a(\omega, x') - E_a(x')] dP(\omega).
\]

Observe that \(E_a(x) \in L^2(D)\) and \(V_a(x, x') \in L^2(D) \otimes L^2(D)\).

Consider

\[
\bar{a}(\omega, x) := a(\omega, x) - E_a(x) \in L^2_P(\Omega) \otimes L^2(D).
\]

Let \((s_m)_{m \in \mathbb{N}}\) and \((e_n)_{n \in \mathbb{N}}\) be orthonormal bases for \(L^2_P(\Omega)\) and \(L^2(D)\), respectively. According to our discussion at the beginning of Section 1.1.1, we have the following representation

\[
\bar{a} = \sum_{n \in \mathbb{N}} \sum_{m \in \mathbb{N}} \bar{a}_{n, m} e_n \otimes s_m, \quad (\bar{a}_{n, m}) \in \ell^2(\mathbb{N} \times \mathbb{N}).
\]
Define $\tilde{a}_m := \sum_{n \in \mathbb{N}} \tilde{a}_{n,m} e_n \in L^2(D)$. Now we verify that $C_{\tilde{a}}$ as given in Definition 1.1.3 coincides with $V_{\tilde{a}}$ defined as above. Indeed if we identify $\tilde{a}$ with an element of $L^2(\Omega \times D)$

\[
V_{\tilde{a}}(x, x') = \int_{\Omega} \tilde{a}(\omega, x) \tilde{a}(\omega, x') dP(\omega)
\]

\[
= \int_{\Omega} \left( \sum_{n \in \mathbb{N}} \sum_{m \in \mathbb{N}} \tilde{a}_{n,m} e_n(x) s_m(\omega) \right) \left( \sum_{n' \in \mathbb{N}} \sum_{m' \in \mathbb{N}} \tilde{a}_{n',m'} e_{n'}(x') s_{m'}(\omega) \right) dP(\omega)
\]

\[
= \sum_{n \in \mathbb{N}} \sum_{m \in \mathbb{N}} \sum_{n' \in \mathbb{N}} \sum_{m' \in \mathbb{N}} \tilde{a}_{n,m} \tilde{a}_{n',m'} e_n(x) e_{n'}(x') \int_{\Omega} s_m(\omega) s_{m'}(\omega) dP(\omega)
\]

\[
= \sum_{m \in \mathbb{N}} \left( \sum_{n \in \mathbb{N}} \tilde{a}_{n,m} e_n(x) \right) \left( \sum_{n' \in \mathbb{N}} \tilde{a}_{n',m} e_{n'}(x') \right)
\]

\[
= \sum_{m \in \mathbb{N}} \tilde{a}_m(x) \tilde{a}_m(x') = C_{\tilde{a}}.
\]

Therefore $V_{\tilde{a}}$ is indeed the correlation of $\tilde{a}$.

By Theorem 1.1.5 we can associate to $V_{\tilde{a}}$ a non-negative definite trace class operator $V_{\tilde{a}} : L^2(D) \rightarrow L^2(D)$ taking the form

\[
(V_{\tilde{a}}v)(x) = \sum_{m \in \mathbb{N}} \langle \tilde{a}_m, v \rangle_{L^2(D)} \tilde{a}_m(x)
\]

\[
= \sum_{m \in \mathbb{N}} \left( \int_D \tilde{a}_m(x') v(x') dx' \right) \tilde{a}_m(x)
\]

\[
= \int_D \sum_{m \in \mathbb{N}} \tilde{a}_m(x) \tilde{a}_m(x') v(x') dx'
\]

\[
= \int_D V_{\tilde{a}}(x, x') v(x') dx'.
\]  \hspace{1cm} (1.2.1)

This is called the Carleman operator. In particular $V_{\tilde{a}}$ is compact and has an eigenpairs sequence $(\lambda_m, \phi_m)_{m=1}^{\infty}$ such that

\[
V_{\tilde{a}} \phi_m = \lambda_m \phi_m
\]  \hspace{1cm} (1.2.2)

with the properties that the eigenvalues are non-negative real such that $\lambda_m \downarrow 0$ and the eigenvectors $(\phi_m)_m$ form an orthonormal sequence in $L^2(D)$.

By Corollary 1.1.6 we have a representation for the covariance given by

\[
V_{\tilde{a}}(x, x') = \sum_{m \in \mathbb{N}} \lambda_m \phi_m(x) \phi_m(x').
\]

Finally we have the following equivalent formulation of Theorem 1.1.7:
Corollary 1.2.1. If \( a \in L^2_P(\Omega) \otimes L^2(D) \), then there exists a sequence \((Y_m)_m \subset L^2_P(\Omega)\) such that \( \mathbb{E} Y_m = 0 \), \( \text{Cov}(Y_m, Y_n) = \delta_{mn} \) for all \( m, n \in \mathbb{N} \) and
\[
a(\omega, x) = \mathbb{E}_a(x) + \sum_{m \in \mathbb{N}} \sqrt{\lambda_m} \phi_m(x) Y_m(\omega)
\] (1.2.3)
where \((\lambda_m)_m\) and \((\phi_m)_m\) are the eigenvalues and eigenvectors of the Carleman operator \( V_a \) defined in (1.2.1). Moreover
\[
Y_m(\omega) = \frac{1}{\sqrt{\lambda_m}} \int_D [a(\omega, x) - \mathbb{E}_a(x)] \phi_m(x) dx.
\]

Definition 1.2.2. Decomposition (1.2.3) is called the Karhunen-Loève expansion of the random field \( a \).

In this context the analogue of Theorem 1.1.9 states that

Corollary 1.2.3. If \( a \in L^2_P(\Omega) \otimes L^2(D) \) has the Karhunen-Loève expansion (1.2.3), then for any \( N \in \mathbb{N} \) it holds
\[
\left\| \hat{a} - \sum_{m=1}^N \sqrt{\lambda_m} \phi_m Y_m \right\|_{L^2_P(\Omega) \otimes L^2(D)}^2 = \sum_{m \geq N+1} \lambda_m.
\] (1.2.4)

1.2.1 More on Convergence of the Karhunen-Loève Expansion

Under certain conditions it can be shown that that the convergence of the truncated Karhunen-Loève expansion is in \( L^\infty(\Omega \times D) \). Namely assuming that the sequence \((Y_m)_m\) is uniformly bounded in \( L^\infty(\Omega) \) and \( \sum_m \sqrt{\lambda_m} \| \phi_m \|_{L^\infty(D)} \) converges, we have
\[
\left\| \hat{a} - \sum_{m=1}^N \sqrt{\lambda_m} \phi_m Y_m \right\|_{L^\infty(\Omega \times D)} \leq C \sum_{m \geq N+1} \sqrt{\lambda_m} \| \phi_m \|_{L^\infty(D)}
\] (1.2.5)
where the constant \( C \) is independent of \( N \). We will use this important observation in Section 4.2.1 where we analyse the overall error of the stochastic collocation method. Therefore inspired by the bound (1.2.5), we are interested in studying the eigenvalue decay and point-wise eigenfunction bounds. The results we are going to present are strictly related to the regularity of the covariance \( V_a \) as stated in the following

Definition 1.2.4. Let \( p, q \in [0, \infty) \). A covariance function \( V_a : D \times D \to \mathbb{R} \) is said to be piecewise analytic (respectively smooth and \( H^{p,q} \)) if there exists a finite partition \( D = \{ D_j \}_{j=1}^K \) of \( D \) into simplices \( D_j \) and there exists a finite family \( G = \{ G_j \}_{j=1}^K \) of open sets in \( \mathbb{R}^d \) such that
\[
\overline{D} = \bigcup_{j=1}^K \overline{D_j}, \quad \overline{D_j} \subset G_j, \forall j = 1, ..., K
\]
and such that \( V_a|_{D_j \times D_j} \) has an extension to \( G_j \times G_j \) which is analytic in \( G_j \times G_j \) (respectively smooth in \( G_j \times G_j \) and is in \( (H^p(G_j) \otimes L^2(D)) \cap (H^q(G_{j'}) \otimes L^2(D)) \) for any pair \((j, j')\).
We refer the reader to [14, Section 2.2] for a proof of the following result.

**Proposition 1.2.5.** Let \( V_a \in L^2(D \times D) \) and let \( \lambda_m \) as in (1.2.2).

1) If \( V_a \) is piecewise analytic, then the eigenvalues satisfy
\[
\lambda_m \leq C_1 e^{-C_2 m^{1/d}}, \quad \forall m
\]
for some constants \( C_1, C_2 > 0 \) depending only on \( V_a \).

2) If \( V_a \) is piecewise \( H^{k,0} \), then the eigenvalues satisfy
\[
\lambda_m \leq C_3 m^{-k/d}, \quad \forall m
\]
for a constant \( C_3 > 0 \) depending only on \( V_a \).

3) If \( V_a \) is piecewise smooth, then the eigenvalues satisfy
\[
\lambda_m \leq C_4 m^{-s}, \quad \forall m
\]
for any \( s > 0 \) with a constant \( C_4 > 0 \) depending only on \( V_a \) and \( s \).

4) If \( V_a \) is a Gaussian covariance function taking the form \( V_a(x, x') = \sigma^2 e^{-\frac{|x-x'|^2}{2\text{diam}(D)^2}} \) for \( \sigma, \gamma > 0 \) called standard deviation and correlation length, respectively, then the eigenvalues satisfy
\[
\lambda_m \leq C_5 \frac{\sigma^2 (1/\gamma)^{m^{1/d}}}{\Gamma(0.5m^{1/d})}, \quad \forall m
\]
where \( \Gamma(\cdot) \) denotes the Gamma function and \( C_5 > 0 \) is independent of \( m \).

The proof of the following proposition can be found in [14, Section 2.3].

**Proposition 1.2.6.** Let \( V_a \in L^2(D \times D) \) and let \( \phi_m \) as in (1.2.2).

1) If \( V_a \) is piecewise \( H^{k,0} \), then \( \phi_m \in H^{k}(D_j) \) for all \( D_j \in D \) and for every \( \varepsilon \in (0, k - d/2] \) there exists a constant \( C_6 = C_6(\varepsilon, k) > 0 \) such that
\[
\|\phi_m\|_{L^\infty(D)} \leq C_6 \lambda_m^{-(d/2+\varepsilon)/k}, \quad \forall m.
\]

2) If \( V_a \) is piecewise smooth, for any \( s > 0 \) there exists a constant \( C_7 = C_7(s, d) > 0 \) such that
\[
\|\phi_m\|_{L^\infty(D)} \leq C_7 \lambda_m^{-s}, \quad \forall m.
\]

Combining Propositions 1.2.5 and 1.2.6 we finally get control on the series (1.2.5).

**Corollary 1.2.7.** If \( V_a \) is piecewise \( H^{k,0} \), then \( \phi_m \in H^{k}(D_j) \) for all \( D_j \in D \) and
\[
\sqrt{\lambda_m} \|\phi_m\|_{L^\infty(D)} \leq K m^{-s}
\]
where \( s = \frac{1}{2}(k - d - 2\varepsilon) \) for an arbitrary \( \varepsilon \in (0, \frac{1}{2}(k - d)] \) and the constant \( K \) is independent of \( m \).
2 Problem Setting

2.1 Strong and Weak Formulations

For \( d \in \mathbb{N} \), let \( D \subset \mathbb{R}^d \) be a bounded Lipschitz domain and \((\Omega, \mathcal{F}, P)\) be a complete probability space. Let \( a, f : \Omega \times D \to \mathbb{R} \) be known random functions. Our model problem is an elliptic partial differential equation with random coefficients and it takes the form: \( P \)-a.s.

\[
\begin{aligned}
- \nabla \cdot (a(\omega, x) \nabla u(\omega, x)) &= f(\omega, x), & x \in D \\
u(\omega, x) &= 0, & x \in \partial D
\end{aligned}
\]  

where the gradient operator \( \nabla \) is taken with respect to \( x \). The main goal is to find a random field \( u : \Omega \times D \to \mathbb{R} \) which satisfies the above problem. As we will see, under certain assumptions on \( a \) and \( f \), it is relatively easy to show that such solution exists and it is unique. On the other hand no explicit expression of \( u \) is known in general. Thus we aim to approximate numerically the solution to (2.1.1).

Consider the Hilbert space \( L^2_P(\Omega) \) of square integrable functions with respect to \( P \) with the usual norm and the Hilbert space \( H^1_0(D) \) of \( H^1(D) \)-functions vanishing at the boundary of \( D \) in a trace sense, equipped with the norm

\[
\|\phi\|_{H^1_0(D)}^2 = \int_D |\nabla \phi(x)|^2 dx.
\]

In this framework we deal with stochastic functions living on a domain which is the Cartesian product of spatial and probabilistic domains. For this reason the choice of

(A1) \( a \in L^2_P(\Omega) \otimes L^2(D) \) with mean and covariance function defined as:

\[
E_a(x) := \int_{\Omega} a(\omega, x) dP(\omega), \quad V_a(x, x') := \int_{\Omega} a(\omega, x) a(\omega, x') dP(\omega) - E_a(x) E_a(x').
\]

(A2) \( a \) is uniformly bounded from below:

\[
\exists a_{\min} > 0 : P(\omega \in \Omega : a(\omega, x) > a_{\min} \forall x \in \overline{D}) = 1.
\]

(A3) \( f \in L^2_P(\Omega) \otimes L^2(D) \) with mean and covariance function defined as:

\[
E_f(x) := \int_{\Omega} f(\omega, x) dP(\omega), \quad V_f(x, x') := \int_{\Omega} f(\omega, x) f(\omega, x') dP(\omega) - E_f(x) E_f(x').
\]
tensor product spaces is natural. In particular we focus on the tensor product Hilbert space \( V_P := L^2_P(\Omega) \otimes H^1_0(D) \) endowed with the inner product
\[
\langle v, w \rangle_{V_P} = \mathbb{E} \int_D \nabla v(\omega, x) \cdot \nabla w(\omega, x) \, dx.
\]

Moreover we introduce the subspace \( V_{P,a} := \{ v \in V_P : \mathbb{E} \int_D a(\omega, x)|\nabla v(\omega, x)|^2 \, dx < \infty \} \) with the norm \( \| v \|_{V_{P,a}}^2 = \mathbb{E} \int_D a(\omega, x)|\nabla v(\omega, x)|^2 \, dx \). Observe that due to (A2) we have the continuous embedding \( V_{P,a} \hookrightarrow V_P \) with
\[
\| v \|_{V_P} \leq \frac{1}{\sqrt{a_{\min}}} \| v \|_{V_{P,a}}.
\]

Define the bilinear form \( B : V_P \times V_P \to \mathbb{R} \) and the linear functional \( F : V_P \to \mathbb{R} \) as follows
\[
B(w, v) := \mathbb{E} \int_D a(\omega, x) \nabla w(\omega, x) \cdot \nabla v(\omega, x) \, dx, \quad F(v) := \mathbb{E} \int_D f(\omega, x)v(\omega, x) \, dx.
\]
We can now reformulate problem (2.1.1) in a variational form: find \( u \in V_P \) such that
\[
B(u, v) = F(v), \quad \forall v \in V_P.
\]

We show in a moment that problem (2.1.3) is well-posed in the sense that a solution exists and it is unique. More precisely if we can prove that \( B \) is continuous and coercive and \( F \) is continuous, then by the Lax-Milgram Theorem we have the claim (see [9], Theorem 2.7.7). In the proof of this result we will use Poincaré’s inequality:
\[
\| v \|_{L^2(D)} \leq C_P \| \nabla v \|_{L^2(D)}, \quad \forall v \in H^1_0(D).
\]

**Lemma 2.1.1.** Under assumptions (A2) and (A3), problem (2.1.3) admits a unique solution \( u \in V_P \) such that
\[
\| u \|_{V_P} \leq \frac{C_P}{a_{\min}} \| f \|_{L^2_P(\Omega) \otimes L^2(D)}.
\]

**Proof.** Holder’s inequality applied twice gives that
\[
|B(u, v)| \leq \| u \|_{V_{P,a}} \| v \|_{V_{P,a}}
\]
and therefore \( B \) is continuous with continuity constant equal to 1. It is straightforward to show also that
\[
B(v, v) = \| v \|^2_{V_{P,a}}.
\]
which in turn gives coerciveness with constant equal to 1.

For what concerns the continuity of $F$ we have

$$|F(v)| = |\mathbb{E}\int_D f v d\mathbf{x}| \leq \mathbb{E}\int_D |f v| d\mathbf{x} \leq \mathbb{E}\|f\|_{L^2(D)}\|v\|_{L^2(D)}$$

$$\leq C_P \mathbb{E}\|f\|_{L^2(D)}\|
abla v\|_{L^2(D)} \leq \frac{C_P}{\alpha_{\min}} \mathbb{E}\|f\|_{L^2(D)} \|
abla v\|_{L^2(D)}$$

$$\leq \frac{C_P}{\sqrt{\alpha_{\min}}} (\mathbb{E}\|f\|_{L^2(D)}^2)^{1/2}\|v\|_{V_{P,a}}$$

$$= \frac{C_P}{\sqrt{\alpha_{\min}}} \|f\|_{L^2_p(\Omega) \otimes L^2(D)} \|v\|_{V_{P,a}}$$

where we have used Holder’s inequality again and Poincaré’s inequality (2.1.4). Thus, by assumption (A3), $F$ is continuous with continuity constant $\frac{C_P}{\sqrt{\alpha_{\min}}} \|f\|_{L^2_p(\Omega) \otimes L^2(D)}$.

Having shown continuity of $B$ and $F$ and coerciveness of $B$, the first part of the lemma is then proved by applying the Lax-Milgram Theorem and the fact that $V_{P,a} \hookrightarrow V_P$.

It remains to show that the estimate in the statement holds. We have

$$\|u\|_{V_P} \leq \frac{1}{a_{\min}} \left( \int_D \mathbb{E}[a^2 |\nabla u|^2] d\mathbf{x} \right)^{1/2} \leq \frac{C_P}{a_{\min}} \left( \int_D \mathbb{E}[|\nabla \cdot (a \nabla u)|^2] d\mathbf{x} \right)^{1/2}$$

$$= \frac{C_P}{a_{\min}} \left( \int_D \mathbb{E}[f^2] d\mathbf{x} \right)^{1/2} = \frac{C_P}{a_{\min}} \|f\|_{L^2_p(\Omega) \otimes L^2(D)}$$

where the first inequality comes from (A2) and the second one from (2.1.4).

\[\square\]

### 2.2 Finite-Dimensional Stochastic Space

In order to deal with the infinite-dimensional problem (2.1.3) we aim to replace the coefficients by finite-dimensional counterparts which can be treated numerically. More precisely we consider a new problem whose coefficients depend only on a finite number $N$ of random variables. We have shown in Chapter 1 that we can decompose a random field with the Karhunen-Loève expansion and we can consider a finite-dimensional truncation of this last one. Hence we choose the new coefficient

$$a_N(\omega, \mathbf{x}) = \mathbb{E}_a(\mathbf{x}) + \sum_{m=1}^N \sqrt{\lambda_m} \phi_m(\mathbf{x}) Y_m(\omega)$$

which converges to $a(\omega, \mathbf{x})$ in the space $L^2_p(\Omega) \otimes L^2(D)$. A similar result holds for $f$. We emphasize the dependence on the random variables $Y_m$’s introducing the notation

$$a_N(\omega, \mathbf{x}) = a_N(Y_1(\omega), ..., Y_N(\omega), \mathbf{x}), \quad f_N(\omega, \mathbf{x}) = f_N(Y_1(\omega), ..., Y_N(\omega), \mathbf{x})$$

where $(Y_m)_{m=1}^N$ are real valued with $\mathbb{E}Y_m = 0$ and $\text{Cov}(Y_m, Y_n) = \delta_{mn}$ (see Corollary 4.1.4). The new problem is then to find $u_N \in V_P$ such that

$$\mathbb{E}\int_D a_N(\omega, \mathbf{x}) \nabla u_N(\omega, \mathbf{x}) \cdot \nabla v(\omega, \mathbf{x}) d\mathbf{x} = \mathbb{E}\int_D f_N(\omega, \mathbf{x}) v(\omega, \mathbf{x}) d\mathbf{x}, \quad \forall v \in V_P. \quad (2.2.1)$$
Observe that by the Doob-Dynkin Lemma (see [3], Theorem 20.1) it holds true that also the function \( u_N \) depends on the same random variables, i.e.

\[
u_N(\omega, x) = u_N(Y_1(\omega), ..., Y_N(\omega), x).
\]

With this characterization the infinite-dimensional probability space \( \Omega \) has been substituted by an \( N \)-dimensional one. We underline the fact that \( a_N \) and \( f_N \) are inexact representations of the coefficients appearing in (2.1.3) and therefore the solution \( u_N \) will also be an approximation of the exact solution \( u \) and the truncation error \( u - u_N \) has to be estimated (see Section 4.2.1).

We will adopt the following notation. Let \( \Gamma_m := Y_m(\Omega) \subset \mathbb{R} \) which can be either bounded or unbounded. Define

\[
\Gamma := \prod_{m=1}^{N} \Gamma_m.
\]

Moreover let assume that the random variables \( Y_1, ..., Y_N \) have joint probability density function \( \rho : \Gamma \rightarrow \mathbb{R}_+ \).

We now consider the tensor product space

\[
V_\rho = L_\rho^2(\Gamma) \otimes H^1_0(D).
\]

Introducing the new notation we can reformulate problem (2.2.1) as follows: find \( u_N \in V_\rho \) such that

\[
\int_{\Gamma} \int_{D} a_N(y, x) \nabla u_N(y, x) \cdot \nabla v_N(y, x) \rho(y) dx dy = \int_{\Gamma} \int_{D} f_N(y, x) v_N(y, x) \rho(y) dx dy, \quad \forall v_N \in V_\rho.
\]  

(2.2.2)

We can equivalently consider \( u_N, a_N \) and \( f_N \) as functions

\[
u_N, a_N, f_N : \Gamma \rightarrow H^1_0(D).
\]

The analogue of problem (2.2.2) is then find \( u_N \in V_\rho \) such that

\[
\int_{D} a_N(y) \nabla u_N(y) \cdot \nabla \psi(x) dx = \int_{D} f_N(y) \psi(x) dx, \quad \forall \psi \in H^1_0(D), \rho\text{-a.e. in } \Gamma.
\]

(2.2.3)

**Remark 1.** The use of the finite-dimensional truncation of the Karhunen-Loève expansion has turned the stochastic problem (2.1.3) into the deterministic parametric elliptic problem (2.2.3) with an \( N \)-dimensional parameter.
3 The Stochastic Collocation Method

The stochastic collocation method aims to approximate numerically the solution $u_N$ to problem (2.2.3). We follow [11] in the description of the method. It is based on standard finite element approximations in the space $H^1_0(D)$ where $D$ is a bounded Lipschitz domain and a collocation, in the space $L^2_\rho(\Gamma)$, on a tensor grid built upon the zeros of polynomials orthogonal with respect to the joint probability density function of the random variables $Y_1, \ldots, Y_N$.

3.1 Full Tensor Grid of Collocation Points

We seek an approximation in the finite-dimensional tensor space $V_{p,h} := \mathcal{P}_p(\Gamma) \otimes S^q_{h}(D)$ where:

- $S^q_{h}(D) \subset H^1_0(D)$ is a finite element space of dimension $N_h$ containing piecewise polynomials of degree $q$ on a uniform triangulation $T_h$ with mesh size $h$.
- $\mathcal{P}_p(\Gamma) \subset L^2_\rho(\Gamma)$ is the span of tensor product polynomials with degree $p = (p_1, \ldots, p_N)$, i.e. $\mathcal{P}_p(\Gamma) := \bigotimes_{m=1}^N \mathcal{P}_{p_m}(\Gamma_m)$ where for each $m \in \{1, \ldots, N\}$

  \[ \mathcal{P}_{p_m}(\Gamma_m) := \text{span}\{y_m^i : i = 0, \ldots, p_m\}. \]

  Hence $N_p := \dim(\mathcal{P}_p(\Gamma)) = \prod_{m=1}^N (p_m + 1)$.

The first step in the approximating process consists in choosing a set of collocation points and applying Lagrange interpolation to $u_N$ at those points. One way to do so is by using a full tensor grid interpolation operator. The evaluation points are chosen to be the roots of polynomials which are orthogonal with respect to a certain auxiliary density function. Standard probability density functions, such as Gaussian or uniform, lead to well known Gauss quadrature nodes and weights. Recall that in general the random variables $Y_m$’s are uncorrelated but not independent and therefore the joint probability density function $\rho$ does not factorize. In this case we introduce $\hat{\rho} : \Gamma \rightarrow \mathbb{R}_+$ such that

\[ \hat{\rho}(y) = \prod_{m=1}^N \hat{\rho}_m(y_m), \quad \left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma)} < \infty. \quad (3.1.1) \]

See Section 4.1 for conditions on the existence of $\hat{\rho}$.

For each $m \in \{1, \ldots, N\}$, we denote by $\{y_{m,k_m}\}_{k_m=1}^{p_m+1} \subset \Gamma_m^{p_m+1}$ the roots of the non-trivial
polynomial \( q_{p+1} \in \mathcal{P}_{p+1}(\Gamma_m) \) such that for any \( v \in \mathcal{P}_p(\Gamma_m) \)

\[
\int_{\Gamma_m} q_{p+1}(y_m)v(y_m)\rho_m(y_m) = 0.
\]

We consider the tensorized grid of collocation points

\[
\mathcal{Y} := \{ \mathbf{y}_k = (y_{m,k_m})_{m=1}^N : 1 \leq k_m \leq p_m + 1 \}.
\]

We order this set introducing the following index associated to the vector \( \mathbf{k} = (k_1, ..., k_N) : \)

\[
k := k_1 + \sum_{i=1}^{N-1} (k_{i+1} - 1) \prod_{j \leq i} (p_j + 1).
\]

More precisely we introduce the bijection

\[
\Psi : \{ \mathbf{k} = (k_1, ..., k_N) : k_m \in \{1, ..., p_m + 1\} \} \longrightarrow \{1, 2, ..., N_p\}, \quad \Psi(\mathbf{k}) = k.
\]

Then we denote by \( \mathbf{y}_k \) the collocation point \( \mathbf{y}_k = (y_{1,k_1}, ..., y_{N,k_N}) \).

We define the full tensor Lagrange interpolant \( \mathcal{I}^{\text{full}} : C^0(\Gamma; H^1_0(D)) \rightarrow \mathcal{P}_p(\Gamma) \otimes H^1_0(D) \),

\[
\mathcal{I}^{\text{full}} v(\mathbf{y}, \mathbf{x}) = \sum_{\mathbf{y}_k \in \mathcal{Y}} v(\mathbf{y}_k, \mathbf{x}) \ell_k(\mathbf{y})
\]

where \( \ell_k(\mathbf{y}) = \prod_{m=1}^{N} \ell_{m,k_m}(y_m) \) and \( \ell_{m,k_m} \in \mathcal{P}_{p_m}(\Gamma_m), \ell_{m,k_m}(y_m) = \prod_{i=1}^{p_m+1} \frac{y_m - y_{m,i}}{y_{m,k_m} - y_{m,i}}. \)

Note that \( \ell_{m,j}(y_{m,i}) = \delta_{ji} \) for \( j, i = 1, ..., p_m + 1 \) and consequently \( \ell_k(y_1) = \delta_{kj} \).

Equivalently by using the global index notation we get

\[
\mathcal{I}^{\text{full}} v(\mathbf{y}, \mathbf{x}) = \sum_{k=1}^{N_p} v(\mathbf{y}_k, \mathbf{x}) \ell_k(\mathbf{y})
\] (3.1.2)

where \( \ell_k(\mathbf{y}) = \prod_{m=1}^{N} \ell_{m,\Psi^{-1}(k)_m}(y_m). \)

By defining one-dimensional interpolants \( \mathcal{I}_{p_m} : C^0(\Gamma_m; H^1_0(D)) \rightarrow \mathcal{P}_{p_m}(\Gamma_m) \otimes H^1_0(D) \),

\[
\mathcal{I}_{p_m} v(y_m, \mathbf{x}) = \sum_{k_m=1}^{p_m+1} v(y_{m,k_m}, \mathbf{x}) \ell_{m,k_m}(y_m)
\] (3.1.3)

we have the equivalent formulation \( \mathcal{I}^{\text{full}} = \bigotimes_{m=1}^{N} \mathcal{I}_{p_m} \).

The semi-discrete approximation in the stochastic domain is obtained by applying the operator \( \mathcal{I}^{\text{full}} \) to \( u_N \), the solution to problem (2.2.3):

\[
u_{N,p} = \mathcal{I}^{\text{full}} u_N \in \mathcal{P}_p(\Gamma) \otimes H^1_0(D).
\]
The second step in the approximating procedure consists in projecting the semi-approximation \( u_{N,p} \) onto the finite element space \( S^q_h(D) \). We may do this using the Galerkin method and we get the final approximation \( u_{N,p,h} : \mathcal{P}_p(\Gamma) \rightarrow S^q_h(D) \) satisfying

\[
\int_D a_N(y) \nabla u_{N,p,h}(y) \cdot \nabla \psi_h(x) dx = \int_D f_N(y) \psi_h(x) dx \quad \forall \psi_h \in S^q_h(D), \forall y \in \mathcal{Y}.
\]

The main goal of the implementations in Chapter 5 will be to approximate the statistics of the solution \( u \) to problem (2.1.3). Therefore we explain how to recover the expected value of the final approximation \( u_{N,p,h} \). We define the Gauss quadrature weights

\[
\omega_{m,k} := \int_{\Gamma_m} \ell_{m,k}^2(y_m) \hat{\rho}_m(y_m) dy_m, \quad \omega_k = \prod_{m=1}^N \omega_{m,\psi^{-1}(k)_m}.
\]

Given these weights, we introduce the following Gauss quadrature formula (see [11], (2.3))

\[
\mathbb{E}_{\hat{\rho}}(u_{N,p,h})(x) = \mathbb{E}_{\hat{\rho}} \left[ I_{\text{full}}^{u_{N,h}}(y,x) \right] = \mathbb{E}_{\hat{\rho}} \left[ \sum_{k=1}^{N_p} u_{N,h}(y_k, x, \xi_{k}(y)) \right] = \sum_{k=1}^{N_p} u_{N,h}(y_k, x) \omega_k
\]

where \( u_{N,h}(y_k, x) \in S^q_h(D) \) indicates the finite element solution of the problem with coefficients evaluated at point \( y_k \).

**Remark 2.** The stochastic collocation method is equivalent to solve \( N_p \) deterministic problems. Note that each of these problems is naturally decoupled. On the other hand \( N_p \) grows exponentially in the number \( N \) of random variables giving place to a huge computational work, the so-called curse of dimensionality.

### 3.2 Sparse Tensor Grid of Collocation Points

We propose an alternative to the approach described in the previous section based on a different choice of the collocation points. We construct the grid by the Smolyak algorithm as described in [11] and [12]. The main goal is to keep the number of points moderate.

Let \( i \in \mathbb{N}^+ \) be a positive integer denoting the level of approximation and \( t : \mathbb{N}_+ \rightarrow \mathbb{N}_+ \) an increasing function denoting the number of collocation points used to build the approximation at level \( i \) with \( t(1) = 1 \). In this context for \( m = 1, \ldots, N \) the one-dimensional interpolation operator \( \mathcal{I}_m^{(i)} : C^0(\Gamma_m; H^1_0(D)) \rightarrow \mathcal{P}_{t(i)-1}(\Gamma_m) \otimes H^1_0(D) \) takes the form

\[
\mathcal{I}_m^{(i)} v(y_m, x) = \sum_{k_m=1}^{t(i)} v(y_{m,k_m}, x) \ell_{m,k_m}(y_m).
\]
Here similarly as before $\{y_{m,k}\}_{k=1}^{t(i)}$ are the roots of the polynomial $q_{m,t(i)} \in \mathcal{P}_{t(i)}(\Gamma_m)$ orthogonal to $\mathcal{P}_{t(i)-1}(\Gamma_m)$ with respect to $\hat{\rho}_m$.

We introduce the difference operators

$$\Delta_{t}^{m} = \begin{cases} \mathcal{I}^{m}_{t} - \mathcal{I}^{m}_{t(i)-1} & , \ i \geq 2 \\ \mathcal{I}^{m}_{t} & , \ i = 1. \end{cases}$$

Given a multi-index $i = (i_1,...,i_N) \in \mathbb{N}_+^N$ we consider a function $g : \mathbb{N}_+^N \rightarrow \mathbb{N}$ strictly increasing in each argument. Let $w \in \mathbb{N}$. Then the isotropic sparse grid approximation $u_{N,p,h}$ is obtained by projecting onto the finite element space $S^h_{N}(D)$ the semi-approximation $u_{N,p} = S^{l,g}_{w} u_{N}$ where

$$S^{l,g}_{w} = \sum_{i \in \mathbb{N}_+^N} g(i) \leq w \prod_{m=1}^{N} \Delta_{t}^{l(i_m)}.$$  \hspace{1cm} (3.2.2)

For $|j| = \sum_{m=1}^{N} j_m$, the following equivalent formulation can be proved by induction:

$$S^{l,g}_{w} = \sum_{i \in \mathbb{N}_+^N} \sum_{j \in \{0,1\}^N} (-1)^{|j|} \prod_{m=1}^{N} \mathcal{I}^{l(i_m)}.$$  \hspace{1cm} (3.2.3)

Therefore the sparse grid approximation can be seen as a linear combination of full tensor product interpolations and the sparse grid $\mathcal{Y}_{\text{sparse}} \subset \Gamma$ is obtained as a superposition of all full tensor grids used in (3.2.3) which correspond to non-zero coefficients. Namely

$$\mathcal{Y}_{\text{sparse}} = \bigcup_{i \in \mathbb{N}_+^N} \bigcup_{j \in \{0,1\}^N} (\mathcal{Y}_{t(i_1)} \times \cdots \times \mathcal{Y}_{t(i_N)})$$  \hspace{1cm} (3.2.4)

where $\mathcal{Y}_{t(i_m)} \subset \Gamma_m$ denotes the grid of points used by $\mathcal{I}^{t(i_m)}$. Recall that the full tensor product grid is obtained as

$$\mathcal{Y} = \mathcal{Y}_{p_1} \times \cdots \times \mathcal{Y}_{p_N}$$

with $\mathcal{Y}_{p_m} \subset \Gamma_m$ the grid of points used by $\mathcal{I}_{p_m}$. So the choice of $w$ and $g$ in the sparse construction is driven by the idea to reduce the number of non-zero terms in (3.2.4).

Figure 3.2.1 shows the significant difference between a full grid and a sparse grid.

The good effectiveness of the sparse collocation method strongly relies on the proper selection of the functions $t$ and $g$. A typical choice which leads to the isotropic Smolyak algorithm is

$$t(i) = \begin{cases} 1, & , \ i = 1 \\ 2^{i-1} + 1, & , \ i \geq 2 \end{cases}, \ g(i) = \sum_{m=1}^{N} (i_m - 1).$$  \hspace{1cm} (3.2.5)
On the other hand it is interesting to notice that the full tensor product method is recovered when we consider

\[ t(i) = i, \quad g(i) = \max_{1 \leq m \leq N} (i_m - 1). \]

The advantage of the sparse tensor product grid over the full tensor product grid can already be seen for a 2-dimensional stochastic domain (see Figure 3.2.2). Since in general the function \( t \) is non surjective we define a left-inverse given by

\[ t^{-1}(k) := \min\{i \in \mathbb{N}_+: t(i) \geq k\}. \]

Note that \( t^{-1}(t(i)) = i \) and \( t(t^{-1}(k)) \geq k \). Let \( t(i) = (t(i_1), ..., t(i_N)) \) and consider the following set

\[ \Theta = \{ p \in \mathbb{N}^N : g(t^{-1}(p + 1)) \leq w \}. \]

It is not hard to see that the Smolyak functions (3.2.5) give rise to

\[
\begin{align*}
\Theta = \left\{ p \in \mathbb{N}^N : \sum_{m=1}^N f(p_m) \leq w \right\}, & \quad f(p_m) = \begin{cases} 
0, & p_m = 0 \\
1, & p_m = 1 \\
\lceil \log_2(p_m) \rceil, & p_m \geq 2 
\end{cases} \\
\end{align*}
\]

Define the polynomial space

\[ \mathcal{P}_\Theta(\Gamma) := \text{span} \left\{ \prod_{m=1}^N y_{p_m} : p = (p_1, ..., p_N) \in \Theta \right\}. \]

Then

\[ S_{w}^{t,g} u_N \in \mathcal{P}_\Theta(\Gamma) \otimes H_0^1(D). \]
Similarly to the previous section, we describe how to compute the first moment of the final approximation $u_{N,p,h}$. For the Smolyak algorithm, it can be shown (see [12], formula (3.9)) that

$$
E_{\hat{\rho}}(u_{N,p,h}^i(x)) = \sum_{i \in \mathbb{N}^N, |i| \leq w} (-1)^{w+N-|i|} \binom{N}{w} E_{\hat{\rho}} \left( \bigotimes_{m=1}^{N} T^{(i_m)}_{m} u_{N,h}^{i_m}(y_m,x) \right)
$$

(3.2.7)

and from (3.1.5)

$$
E_{\hat{\rho}} \left( \bigotimes_{m=1}^{N} T^{(i_m)}_{m} u_{N,h}^{i_m}(y_m,x) \right) = \sum_{k=1}^{N_{(i)}} u_{N,h}^{i_k}(y_k,x) \omega_k
$$

where $\omega_k$ are the same as in (3.1.4) and again $u_{N,h}^{i_k}(y_k,x)$ is the finite element solution of the problem with coefficients evaluated at point $y_k$.

### 3.2.1 Anisotropic Version

It is possible to construct an even refined algorithm acting differently on each direction $y_m$. This may be useful in situations where the convergence rate is poor in some
directions with respect to others. This anisotropy can be described introducing some weights \( \alpha = (\alpha_1, ..., \alpha_N) \) in the function \( g \). One possible choice can be for instance

\[
g(i; \alpha) := \sum_{m=1}^{N} \frac{\alpha_m}{\alpha_{\min}} (i_m - 1), \quad \alpha_{\min} := \min_{1 \leq m \leq N} \alpha_m.
\]

Consequently the corresponding anisotropic sparse interpolation operator takes the form

\[
S^\ell_{w, \alpha} = \sum_{i \in \mathbb{N}^N} \prod_{m=1}^{N} \Delta_{\ell(i_m)}. \quad \text{if } g(i\alpha) \leq w
\]

Observe that the isotropic Smolyak method described in the previous section is a special case of the anisotropic formula when we take all the components of the weight vector to be equal, i.e. \( \alpha_1 = ... = \alpha_N \).

We will see in the next chapter (cf. Section 4.4) how the selection of weights is related to the analytic dependence of the solution \( u_N \) with respect to each of the random variables \( Y_m \). The key idea is to place more points in those directions where the convergence rate in the random domain is slower.
4 Convergence of the Method

Before studying the convergence of the stochastic collocation method we need to impose some more requirements on the data of the problem which in turn, as we will see, imply regularity of the stochastic behaviour of the solution $u_N$ to problem

$$
\int_D a_N(y) \nabla u_N(y) \cdot \nabla \psi(x) dx = \int_D f_N(y) \psi(x) dx, \quad \forall \psi \in H_0^1(D), \rho\text{-a.e. in } \Gamma. \quad (4.0.1)
$$

The results presented in this chapter can be partially found in [11]. In the following section for convenience we drop the subscript $N$ which indicates the finite-dimensional noise dependence.

4.1 Regularity Results

We introduce the following weight $\sigma : \Gamma \to \mathbb{R}^+$ which allows to keep control on the exponential growth at infinity of some function:

$$
\sigma(y) = \prod_{m=1}^N \sigma_m(y_m), \quad \sigma_m(y_m) = \begin{cases} 1, & \Gamma_m \text{ is bounded} \\ e^{-\alpha_m|y_m|}, & \Gamma_m \text{ is unbounded} \end{cases}
$$

for some $\alpha_m > 0$. We consider the corresponding functional space

$$
C_\sigma^0(\Gamma; V) = \left\{ v : \Gamma \to V, v \text{ continuous in } y, \max_{y \in \Gamma} \sigma(y)\|v(y)\|_V < \infty \right\}
$$

where $V$ is an Hilbert space defined on $D$. As we have announced at the beginning of the chapter, we make the following assumptions on the data of problem (2.2.3):

(A4) $f \in C_\sigma^0(\Gamma; L^2(D))$.

(A5) The joint probability density function $\rho$ is such that $\forall y \in \Gamma$,

$$
\rho(y) \leq C_\rho e^{-\sum_{m=1}^N (\delta_m y_m)^2} \quad (4.1.1)
$$

for some $C_\rho > 0$ and $\delta_m = 0$ in the case $\Gamma_m$ is bounded and $\delta_m > 0$ when $\Gamma_m$ is unbounded.

We have remarked in Section 3.1 that we need to introduce an auxiliary probability density function satisfying (3.1.1). The following condition is sufficient

$$
C_{\min}^m e^{-(\delta_m y_m)^2} \leq \hat{\rho}_m(y_m) < C_{\max}^m e^{-(\delta_m y_m)^2}, \quad \forall y_m \in \Gamma_m
$$
for some constants $c^{(m)}_{\min}, c^{(m)}_{\max} > 0$ which do not depend on $y_m$. Then (3.1.1) holds with

$$\left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma)} \leq \frac{C_\rho}{c^{(m)}_{\min}}, \quad c^{(m)}_{\min} := \prod_{m=1}^N c^{(m)}_{\min}.$$

Under the previous assumptions the following continuous embeddings hold true:

$$C_\sigma^0(\Gamma; V) \hookrightarrow L^2_\rho(\Gamma; V) \hookrightarrow L^2_\rho(\Gamma; V).$$

Indeed we have

$$\|v\|_{L^2_\rho(\Gamma; V)} \leq \left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma)} \|v\|_{L^2_\rho(\Gamma; V)} \leq \left( \frac{C_\rho}{c^{(m)}_{\min}} \right)^{1/2} \|v\|_{L^2_\rho(\Gamma; V)}$$

proving continuity of the second embedding. For the first one we have

$$\|v\|_{L^2_\rho(\Gamma; V)}^2 = \int_{\Gamma} (\sigma(y)\|v(y)\|^2) \frac{\hat{\rho}(y)}{\sigma^2(y)} dy \leq \|v\|_{C_\sigma^2(\Gamma; V)}^2 \int_{\Gamma} \frac{\hat{\rho}(y)}{\sigma^2(y)} dy = \|v\|_{C_\sigma^2(\Gamma; V)}^2 \prod_{m=1}^N \int_{\Gamma_m} \frac{\hat{\rho}_m(y_m)}{\sigma_m^2(y_m)} dy_m.$$

Denote with $M_m := \int_{\Gamma_m} \hat{\rho}_m(y_m) dy_m$. We have two possible cases. If $\Gamma_m$ is bounded then $\sigma_m(y_m) = 1$ and $\delta_m = 0$ and therefore

$$M_m = \int_{\Gamma_m} \hat{\rho}_m(y_m) dy_m \leq \int_{\Gamma_m} C^{(m)}_{\max} e^{-\alpha_m y_m^2} dy_m = C^{(m)}_{\max} |\Gamma_m|$$

where $|\Gamma_m|$ denotes the volume of $\Gamma_m$. On the other hand if $\Gamma_m$ is unbounded then $\sigma_m(y_m) = e^{-\alpha_m |y_m|}$ and $\delta_m > 0$. Thus

$$M_m = \int_{\Gamma_m} \hat{\rho}_m(y_m) e^{2\alpha_m |y_m|} dy_m = \int_{\Gamma_m} \hat{\rho}_m(y_m) \left( e^{-\delta_m^2 y_m^2 + 2\alpha_m |y_m|} \right) dy_m \leq C^{(m)}_{\max} \int_{\Gamma_m} \left( e^{-\delta_m^2 y_m^2 + 2\alpha_m |y_m|} \right) dy_m \leq C^{(m)}_{\max} \int_{\Gamma_m} e^{-\delta_m^2 y_m^2 + \frac{2\alpha_m |y_m|}{\sigma_m^2}} dy_m \leq C^{(m)}_{\max} \frac{\sqrt{2\pi}}{\delta_m} e^{\frac{-(\sigma_m^2)^2}{4\delta_m^2}}.$$

Hence $C_\sigma^0(\Gamma; V)$ is continuously embedded in $L^2_\rho(\Gamma; V)$ for either $\Gamma_m$ bounded or unbounded.

**Lemma 4.1.1.** If $f \in C_\sigma^0(\Gamma; L^2(D))$ and $a$ is uniformly bounded from below by $a_{\min} > 0$, then the solution to (4.0.1) is such that $u \in C_\sigma^0(\Gamma; H_0^1(D))$.
Proof. By the definition of the functional space \( C_{\sigma}^0(\Gamma; H_0^1(D)) \) we have

\[
\|u\|_{C_{\sigma}^0(\Gamma; H_0^1(D))} = \sup_{y \in \Gamma} \sigma(y) \|u(y)\|_{H_0^1(D)} = \sup_{y \in \Gamma} \sigma(y) \|\nabla u(y)\|_{L^2(D)} \\
\leq \frac{1}{a_{\min}} \sup_{y \in \Gamma} \sigma(y) \|a(y)\nabla u(y)\|_{L^2(D)} \\
\leq \frac{C_P}{a_{\min}} \sup_{y \in \Gamma} \sigma(y) \|f(y)\|_{L^2(D)} \\
= \frac{C_P}{a_{\min}} \|f(y)\|_{C_{\sigma}^0(\Gamma; L^2(D))}
\]

where in the second inequality we have used Poincaré’s inequality. \( \square \)

The last result we prove before investigating the convergence of the stochastic collocation method concerns a kind of regularity of the solution \( u \) in the random domain. For later convenience we introduce the following notation.

\[
\Gamma_m^* := \prod_{j=1, j \neq m}^N \Gamma_j, \quad y_m^* := (y_1, \ldots, y_{m-1}, y_{m+1}, \ldots, y_N) \in \Gamma_m^*.
\]

Similarly we write

\[
\hat{\rho}_m^* := \prod_{j=1, j \neq m}^N \hat{\rho}_j, \quad \sigma_m^* := \prod_{j=1, j \neq m}^N \sigma_j.
\]

With slight abuse of notation we write \( v(y, x) = v(y_m, y_m^*, x) \) for any \( m = 1, \ldots, N \).

**Lemma 4.1.2.** Assume that for every \( y \in \Gamma \) and any \( m \in \{1, \ldots, N\} \) there exists \( \gamma_m < \infty \) such that for all \( k \in \mathbb{N} \)

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

Then the solution \( u(y_m, y_m^*, x) \) to problem (4.0.1) as a function \( u : \Gamma_m \to C_{\sigma_m^*}^0(\Gamma_m^*; H_0^1(D)) \) admits an analytic extension \( \tilde{u}(z, y_m^*, x) \) in the region

\[
\Sigma(\Gamma_m; \tau_m) := \{z \in \mathbb{C} : \text{dist}(z, \Gamma_m) \leq \tau_m\} \subset \mathbb{C}
\]

with \( 0 < \tau_m < \frac{1}{2\gamma_m} \). Moreover for all \( z \in \Sigma(\Gamma_m; \tau_m) \)

\[
\sigma_m(|z|) \|\tilde{u}(z)\|_{C_{\sigma_m^*}^0(\Gamma_m^*; H_0^1(D))} \leq \frac{C_p e^{\alpha_m \tau_m}}{2a_{\min}(1 - 2\tau_m \gamma_m)} (1 + 2\|f\|_{C_{\sigma}^0(\Gamma; L^2(D))})
\]

where \( C_P \) is the constant appearing in Poincaré’s inequality (2.1.4).
Proof. Consider the weak problem (4.0.1) which, after suppressing subscripts, takes the form
\[ \int_D a(y) \nabla u(y) \cdot \nabla \psi(x) dx = \int_D f(y) \psi(x) dx, \quad \forall \psi \in H^1_0(D), \rho\text{-a.e. in } \Gamma. \]

Differentiating \( k \) times with respect to \( y_m \) and using Leibniz’s product rule we end up with
\[ \sum_{l=0}^k \binom{k}{l} \int_D (\partial_{y_m}^l a(y)) \nabla \partial_{y_m}^{k-l} u(y) \cdot \nabla \psi(x) dx = \int_D (\partial_{y_m}^k f(y)) \psi(x) dx \]
and reordering terms we obtain
\[ \int_D a(y) \nabla \partial_{y_m}^k u(y) \cdot \nabla \psi(x) dx = \]
\[ - \sum_{l=1}^k \binom{k}{l} \int_D (\partial_{y_m}^l a(y)) \nabla \partial_{y_m}^{k-l} u(y) \cdot \nabla \psi(x) dx + \int_D (\partial_{y_m}^k f(y)) \psi(x) dx. \]

Setting \( \psi = \partial_{y_m}^k u \) and dropping the variable dependence, we get
\[ \int_D a(y) |\nabla \partial_{y_m}^k u| dx = - \sum_{l=1}^k \binom{k}{l} \int_D (\partial_{y_m}^l a(y)) |\nabla \partial_{y_m}^{k-l} u| dx + \int_D (\partial_{y_m}^k f(y)) \partial_{y_m}^k u. \]

Thus
\[ \|\sqrt{a} \nabla \partial_{y_m}^k u\|_2^2 = - \sum_{l=1}^k \binom{k}{l} \int_D \frac{\partial_{y_m}^l a(y)}{a(y)} (\sqrt{a} \nabla \partial_{y_m}^{k-l} u) \cdot (\sqrt{a} \nabla \partial_{y_m}^k u) dx + \int_D (\partial_{y_m}^k f(y)) \partial_{y_m}^k u \]
\[ \leq \sum_{l=1}^k \binom{k}{l} \left\| \frac{\partial_{y_m}^l a(y)}{a(y)} \right\|_{L^\infty(D)} \left\| (\sqrt{a} \nabla \partial_{y_m}^{k-l} u) \cdot (\sqrt{a} \nabla \partial_{y_m}^k u) \right\|_{L^2(D)} + \int_D (\partial_{y_m}^k f(y)) \partial_{y_m}^k u. \]

Applying Holder’s inequality to the terms in absolute value, we have
\[ \|\sqrt{a} \nabla \partial_{y_m}^k u\|_2^2 \leq \sum_{l=1}^k \binom{k}{l} \left\| \frac{\partial_{y_m}^l a(y)}{a(y)} \right\|_{L^\infty(D)} \|\sqrt{a} \nabla \partial_{y_m}^{k-l} u\|_{L^2(D)} \|\sqrt{a} \nabla \partial_{y_m}^k u\|_{L^2(D)} + \int_D (\partial_{y_m}^k f(y)) \partial_{y_m}^k u \]
\[ + \|\partial_{y_m}^k f\|_{L^2(D)} \|\partial_{y_m}^k u\|_{L^2(D)}. \]
Dividing both sides by \( \| \sqrt{a} \nabla \partial^k u \|_{L^2(D)} \) and recalling assumption (A2), it holds
\[
\| \sqrt{a} \nabla \partial^k u \|_{L^2(D)} \leq \sum_{l=1}^{k} \binom{k}{l} \frac{\partial^l u}{a} \| \sqrt{a} \nabla \partial^{k-l} u \|_{L^2(D)} + \frac{\| \partial^k u \|_{L^2(D)}}{\sqrt{a} \min} \| \sqrt{a} \nabla \partial^k u \|_{L^2(D)}.
\]

Application of Poincaré’s inequality gives
\[
\| \sqrt{a} \nabla \partial^k u \|_{L^2(D)} \leq \sum_{l=1}^{k} \binom{k}{l} \frac{\partial^l u}{a} \| \sqrt{a} \nabla \partial^{k-l} u \|_{L^2(D)} + \frac{\| \partial^k u \|_{L^2(D)}}{\sqrt{a} \min} \| \sqrt{a} \nabla \partial^k u \|_{L^2(D)}.
\]

Let define \( R_k := \frac{\| \sqrt{a} \nabla \partial^k u \|_{L^2(D)}}{k!} \). Using the bounds in the assumption of the lemma it holds true that
\[
R_k \leq \sum_{l=1}^{k} \gamma^l_m R_{k-l} + \frac{C_P}{\sqrt{a} \min} \gamma^k_m (1 + \| f \|_{L^2(D)}).
\]

By induction we are going to prove that the following relation holds:
\[
R_k \leq \frac{1}{2} (2\gamma_m)^k \left( R_0 + \frac{C_P}{\sqrt{a} \min} (1 + \| f \|_{L^2(D)}) \right).
\]

For convenience set the constant \( C := \frac{C_P}{\sqrt{a} \min} (1 + \| f \|_{L^2(D)}) \). It is easily verified that for
\(k = 1\), (4.1.5) follows from (4.1.4). Now suppose (4.1.5) holds true for \(k > 1\). Then

\[
R_{k+1} \leq \sum_{l=1}^{k+1} \gamma_l m R_{k+1-l} + \gamma_m^{k+1} C = \sum_{l=1}^{k} \gamma_l m R_{k+1-l} + \gamma_m^{k+1} R_0 + \gamma_m^{k+1} C
\]

\[
\leq \sum_{l=1}^{k} \gamma_l m \frac{1}{2} (2\gamma_m)^{k+1-l} (R_0 + C) + \gamma_m^{k+1} (R_0 + C)
\]

\[
= \frac{1}{2} (2\gamma_m)^{k+1} (R_0 + C) \sum_{l=1}^{k} 2^{-l} + \gamma_m^{k+1} (R_0 + C)
\]

\[
= \frac{1}{2} (2\gamma_m)^{k+1} (R_0 + C) \left(1 - \frac{1}{2^k}\right) + \gamma_m^{k+1} (R_0 + C)
\]

\[
= \frac{1}{2} (2\gamma_m)^{k+1} (R_0 + C).
\]

Observe that (4.1.3) gives for \(k = 0\)

\[
R_0 = \|\sqrt{a} \nabla u\|_{L^2(D)} \leq \frac{C_P}{\sqrt{a_{\min}}} \|f\|_{L^2(D)}.
\]

Moreover

\[
R_k \geq \frac{\sqrt{a_{\min}} \|\nabla \partial_{y_m}^k u\|_{L^2(D)}}{k!}
\]

Therefore we finally achieve

\[
\|\nabla \partial_{y_m}^k u\|_{L^2(D)} \leq \frac{R_k}{\sqrt{a_{\min}}} \leq \frac{1}{\sqrt{a_{\min}}} \frac{1}{2} (2\gamma_m)^k \left( R_0 + \frac{C_P}{\sqrt{a_{\min}}} (1 + \|f\|_{L^2(D)}) \right)
\]

\[
\leq \frac{1}{\sqrt{a_{\min}}} \frac{1}{2} (2\gamma_m)^k \left( \frac{C_P}{\sqrt{a_{\min}}} \|f\|_{L^2(D)} + \frac{C_P}{\sqrt{a_{\min}}} (1 + \|f\|_{L^2(D)}) \right)
\]

\[
= \frac{C_P}{2a_{\min}} (2\gamma_m)^k \left(1 + 2 \|f\|_{L^2(D)} \right).
\]

(4.1.6)

Now fix \(y_m \in \Gamma_m\) and define the power series \(\tilde{u} : \mathbb{C} \to C^0_{\sigma_m} (\Gamma_m; H^1_0(D))\)

\[
\tilde{u}(z, y_m, x) := \sum_{k=0}^{\infty} \frac{(z - y_m)^k}{k!} \partial_{y_m}^k u(y_m, y_m, x).
\]

We aim to prove that the above series converges to the solution \(u\) in a complex disc centred at \(y_m\). For \(z \in \mathbb{C}\) we have

\[
\|\tilde{u}(z)\|_{C^0_{\sigma_m} (\Gamma_m; H^1_0(D))} \leq \sum_{k=0}^{\infty} \frac{(z - y_m)^k}{k!} \|\partial_{y_m}^k u(y_m)\|_{C^0_{\sigma_m} (\Gamma_m; H^1_0(D))}
\]

\[
= \sum_{k=0}^{\infty} \frac{(z - y_m)^k}{k!} \|\nabla \partial_{y_m}^k u(y_m)\|_{C^0_{\sigma_m} (\Gamma_m; L^2(D))}.
\]

32
From (4.1.6) we get for all \( k \in \mathbb{N}_0 \)
\[
\| \nabla \partial_{y_m}^k u(y_m) \|_{C_{\sigma_n}^0 (\Gamma_n^*; L^2(D))} = \max_{y_m \in \Gamma_n^*} \sigma_n^* (y_m^*) \| \nabla \partial_{y_m}^k u(y_m, y_m^*) \|_{L^2(D)} \leq \max_{y_m \in \Gamma_n^*} \sigma_n^* (y_m^*) \frac{C_P}{2a_{\min}} (2\gamma_m)^k k! \left( 1 + 2 \| f(y_m) \|_{L^2(D)} \right) \]
\[
= \frac{C_P}{2a_{\min}} (2\gamma_m)^k k! \left( \max_{y_m \in \Gamma_n^*} \sigma_n^* (y_m^*) + 2 \| f(y_m) \|_{C_{\sigma_n}^0 (\Gamma_n^*; L^2(D))} \right) \leq \frac{C_P}{2a_{\min}} (2\gamma_m)^k k! \left( 1 + 2 \| f(y_m) \|_{C_{\sigma_n}^0 (\Gamma_n^*; L^2(D))} \right)
\]

where the last inequality follows by definition of the weight \( \sigma \). Inserting this bound in the previous display
\[
\| \tilde{u}(z) \|_{C_{\sigma_n}^0 (\Gamma_n^*; H^1_0(D))} \leq \sum_{k=0}^{\infty} \frac{(z - y_m)^k}{k!} \| \nabla \partial_{y_m}^k u(y_m) \|_{C_{\sigma_n}^0 (\Gamma_n^*; L^2(D))} \leq \sum_{k=0}^{\infty} \frac{(z - y_m)^k}{k!} \frac{C_P}{2a_{\min}} (2\gamma_m)^k k! \left( 1 + 2 \| f(y_m) \|_{C_{\sigma_n}^0 (\Gamma_n^*; L^2(D))} \right) \]
\[
= \frac{C_P}{2a_{\min}} \left( 1 + 2 \| f(y_m) \|_{C_{\sigma_n}^0 (\Gamma_n^*; L^2(D))} \right) \sum_{k=0}^{\infty} [(z - y_m) (2\gamma_m)]^k. \tag{4.1.7}
\]

Therefore we can conclude that the power series converges for all \( z \in \mathbb{C} \) such that
\[
\text{dist}(z, y_m) \leq \tau_m < \frac{1}{2\gamma_m}.
\]

This reasoning is independent on the choice of \( y_m \in \Gamma_n^* \). Hence by a continuation argument the series converges in \( \Sigma(\Gamma_n^*, \tau_m) \) for \( \tau_m < \frac{1}{2\gamma_m} \). Finally observe that the power series converges exactly to \( u \). Indeed for all \( z \in \mathbb{C} \)
\[
\left\| u(z) - \sum_{k=0}^{n} \frac{(z - y_m)^k}{k!} \partial_{y_m}^k u(y_m, y_m^*, x) \right\|_{C_{\sigma_n}^0 (\Gamma_n^*; H^1_0(D))} \leq \sup_{\text{dist}(s, y_m) \leq |z - y_m|} \| \partial_{y_m}^{n+1} u(s, y_m^*, x) \|_{C_{\sigma_n}^0 (\Gamma_n^*; H^1_0(D))} \frac{|z - y_m|^{n+1}}{(n+1)!} \leq \frac{C_P}{2a_{\min}} |z - y_m|^{n+1} (2\gamma_m)^n \left( 1 + 2 \| f(y_m) \|_{C_{\sigma_n}^0 (\Gamma_n^*; L^2(D))} \right)
\]
and this upper-bound goes to zero as \( n \to \infty \) whenever \( |z - y_m| < \frac{1}{2\gamma_m} \).

Note that for \( z \in \Sigma(\Gamma_n^*, \tau_m) \) it holds that
\[
\sigma_m (\Re z) \leq e^{\alpha_m \tau_m} \sigma_m (y_m).
\]

33
Along the same line as above, the estimate in the statement follows from
\[
\sigma_m(\mathbb{R}^z)\|u(z)\|_{C_{\gamma_m}^\alpha(\Gamma_m^*:H_0^1(D))} \leq e^{\alpha_m\tau_m}\sigma_m(y_m)\|u(z)\|_{C_{\gamma_m}^\alpha(\Gamma_m^*:H_0^1(D))}
\leq e^{\alpha_m\tau_m}\max_{y_m \in \Gamma_m}\sigma_m(y_m)\frac{C_P}{2a_{\min}}\left(1 + 2\|f\|_{C_{\gamma_m}^\alpha(\Gamma_m^*:L^2(D))}\right)\sum_{k=0}^{\infty}[(z - y_m)(2\gamma_m)]^k
\]
\[
eq e^{\alpha_m\tau_m}\frac{C_P}{2a_{\min}}\left(\max_{y_m \in \Gamma_m}\sigma_m(y_m) + 2\|f\|_{C_{\gamma_m}^\alpha(\Gamma_m^*:L^2(D))}\right)\sum_{k=0}^{\infty}[(z - y_m)(2\gamma_m)]^k
\]
\[
\leq e^{\alpha_m\tau_m}\frac{C_P}{2a_{\min}}(1 + 2\|f\|_{C_{\gamma_m}^\alpha(\Gamma_m^*:L^2(D))})\sum_{k=0}^{\infty}[(z - y_m)(2\gamma_m)]^k
\]
\[
\leq \frac{C_Pe^{\alpha_m\tau_m}}{2a_{\min}(1 - 2\tau_m\gamma_m)}(1 + 2\|f\|_{C_{\gamma_m}^\alpha(\Gamma_m^*:L^2(D))})
\]
where again we have used in the second-to-last inequality the fact that \(\sigma_m \leq 1\) in \(\Gamma_m\).

\[\square\]

**Remark 3.** The coefficients \(a_N\) and \(f_N\) in (4.0.1) are decomposed by using truncations of the Karhunen-Loève expansions of the corresponding infinite-dimensional random fields. In this case assumptions of Lemma 4.1.2 are fulfilled. In particular for
\[
a_N(\omega, x) = \mathbb{E}_a(x) + \sum_{m=1}^{N} \sqrt{\lambda_m}\phi_m(x)Y_m(\omega)
\]
provided that \(a_N(\omega, x) \geq a_{\min}\) in \(D\) and \(P\text{-a.s. in } \Omega\), we have
\[
\left\| \frac{\partial^k_{y_m} a_N(y)}{a_N(y)} \right\|_{L^\infty(D)} \leq \begin{cases} \sqrt{\lambda_m}\|\phi_m\|_{L^\infty(D)}, & k = 1 \\ \frac{\lambda_m}{a_{\min}}, & k > 1 \end{cases}
\]
and we can take \(\gamma_m = \frac{\sqrt{\lambda_m}\|\phi_m\|_{L^\infty(D)}}{a_{\min}}\). Similarly for
\[
f_N(\omega, x) = \mathbb{E}_f(x) + \sum_{m=1}^{N} \sqrt{\mu_m}\phi_m(x)Y_m(\omega)
\]
we have
\[
\left\| \frac{\partial^k_{y_m} f_N(y)}{f_N(y)} \right\|_{L^2(D)} \leq \begin{cases} \sqrt{\mu_m}\|\phi_m\|_{L^2(D)}, & k = 1 \\ \frac{\mu_m}{1 + \|f_N(y)\|_{L^2(D)}}, & k > 1 \end{cases}
\]
and we can take \(\gamma_m = \frac{\sqrt{\mu_m}\|\phi_m\|_{L^2(D)}}{\mu_m}\).

We may also be interested, as we will do in Chapter 5, in a truncated exponential decomposition of the diffusion coefficient. More precisely,
\[
\log[a_N(\omega, x) - a_{\min}] = \mathbb{E}_a(x) + \sum_{m=1}^{N} \sqrt{\lambda_m}\phi_m(x)Y_m(\omega).
\]

34
In this case we have
\[ \left\| \frac{\partial^k_y a_N(y)}{a_N(y)} \right\|_{L^\infty(D)} \leq \left( \sqrt{\lambda_m} \| \phi_m \|_{L^\infty(D)} \right)^k, \quad k \geq 1 \]
and we can take \( \gamma_m = \sqrt{\lambda_m} \| \phi_m \|_{L^\infty(D)}. \)

4.2 Convergence Analysis

The main goal of this section is to provide an a priori estimate on the total error between the exact solution \( u \) of problem (2.1.3) and the approximation \( u_{N,p,h} \) we finally recover by applying the stochastic collocation method. By considering the subsequent approximations we have introduced in our analysis, we can naturally split the total error in the following way
\[ \| u - u_{N,p,h} \|_{V_P} \leq \| u - u_N \|_{V_P} + \| u_N - u_{N,p} \|_{V_P} + \| u_{N,p} - u_{N,p,h} \|_{V_P}, \]
where \( V_P = L^2_P(\Omega) \otimes H^1_0(D) \) as defined in Chapter 2.

4.2.1 Error of the Truncation

We first focus on the term \( \| u - u_N \|_{V_P} \). Here \( u_N \) indicates the exact solution of the weak problem (2.2.3) where the random coefficients of (2.1.3) have been substituted by the corresponding Karhunen-Loève truncated expansions. Thus we aim to give an upper bound for the truncation error of \( u \) depending on the truncation errors of \( a \) and \( f \) which have been investigated in Section 1.2. Consider the two variational formulations of finding \( u, u_N \in V_P \) such that respectively it holds \( \forall \psi \in H^1_0(D) \), \( P \)-a.e. in \( \Omega \)
\[ \int_D a(\omega, x) \nabla u(\omega, x) \cdot \nabla \psi(x) dx = \int_D f(\omega, x) \psi(x) dx, \quad (4.2.1) \]
\[ \int_D a_N(\omega, x) \nabla u_N(\omega, x) \cdot \nabla \psi(x) dx = \int_D f_N(\omega, x) \psi(x) dx. \quad (4.2.2) \]
Now dropping the variable dependence for shortness of the notation, the following holds \( P \)-a.e. in \( \Omega \) and \( \forall \psi \in H^1_0(D) \)
\[ \int_D a \nabla (u - u_N) \cdot \nabla \psi dx = \int_D a \nabla u \cdot \nabla \psi dx - \int_D a_N \nabla u_N \cdot \nabla \psi dx \]
\[ + \int_D a_N \nabla u_N \cdot \nabla \psi dx - \int_D a \nabla u_N \cdot \nabla \psi dx \]
\[ = \int_D (f - f_N) \psi dx + \int_D (a_N - a) \nabla u_N \cdot \nabla \psi dx. \]
Therefore we end up with the $P$-a.e. estimate

$$
\|u - u_N\|_{H^1_0(D)} \leq \frac{1}{\alpha_{\min}} \sup_{\psi \in H^1_0(D)} \frac{\int_D a \nabla (u - u_N) \cdot \nabla \phi}{\|\psi\|_{H^1_0(D)}} \lesssim \|f - f_N\|_{L^2(D)} + \|a - a_N\|_{L^\infty(D)} \|\nabla u_N\|_{L^2(D)}
$$

where the notation $\lesssim c$ means that there exists a constant $K$ such that $b \leq Kc$. Remarkably the constants we have omitted in the previous estimate do not depend on $N$. Finally we obtain

$$
\|u - u_N\|_{V_P} \lesssim \|f - f_N\|_{L^2_0(\Omega) \otimes L^2(D)} + \|a - a_N\|_{L^2_0(\Omega; L^\infty(D))} \|f_N\|_{L^2_0(\Omega) \otimes L^2(D)}.
$$

Given the Karhunen-Loève expansions of $a$ and $f$

$$
a(\omega, x) = E_a(x) + \sum_{m=1}^{\infty} \sqrt{\lambda_m} \phi_m(x) Y_m(\omega), \quad f(\omega, x) = E_f(x) + \sum_{m=1}^{\infty} \sqrt{\mu_m} \varphi_m(x) Y_m(\omega),
$$

we have shown in Section 1.2 that the following estimate holds

$$
\|f - f_N\|_{L^2_0(\Omega) \otimes L^2(D)}^2 = \sum_{m \geq N+1} \mu_m.
$$

On the other hand, assuming that the sequence $(Y_m)_m$ is uniformly bounded in $L^\infty_0(\Omega)$,

$$
\|a - a_N\|_{L^2_0(\Omega; L^\infty(D))} \leq \|a - a_N\|_{L^\infty(\Omega \times D)} \leq C \sum_{m \geq N+1} \sqrt{\lambda_m} \|\phi_m\|_{L^\infty(D)}.
$$

Moreover depending on the regularity, as in Definition 1.2.4, of the covariance functions $V_a$ and $V_f$ of $a$ and $f$ respectively, we have presented explicit bounds for the eigenvalues and eigenfunctions (see Section 1.2.1).

### 4.2.2 Finite Element Error

Here we concentrate on the analysis of the error caused by the finite element approximation as described at the end of Section 3.1. We refer the reader to [9]. We are interested in estimating $\|u_{N,p} - u_{N,p,h}\|_{V_P}$. First of all observe that by passing to the probability space $L^2_\rho(G)$ we have

$$
\|u_{N,p} - u_{N,p,h}\|_{V_P} = \|u_{N,p} - u_{N,p,h}\|_{V_\rho},
$$
where \( V_\rho = L^2_\rho (\Gamma) \otimes H^1_0 (D) \) as defined in Section 2.2. Consider \( V_h(D) \) a finite element space of continuous piecewise polynomials defined on a uniform triangulation \( T_h \) of the domain \( D \subset \mathbb{R}^d \) with maximum mesh size \( h \). The approximation error can be controlled by the mesh size. Namely

\[
\| u_{N,p} - u_{N,p,h} \|_{V_\rho} \leq C \min_{v \in L^2_\rho (\Gamma) \otimes V_h(D)} \| u_{N,p} - v \|_{V_\rho} \leq C (u_{N,p}, n) h^n
\]

where both \( C \) and \( C (u_{N,p}, n) \) are independent of the mesh size and \( n \) is a positive integer determined by the smoothness of \( u_{N,p} \) in \( D \) and the degree of the finite element space. In particular for \( S^q_h(D) \subset H^1_0 (D) \) containing all continuous piecewise polynomials of degree \( q \) we have for \( l \in \{ 0, 1 \} \)

\[
\| u_{N,p} - u_{N,p,h} \|_{L^2_\rho (\Gamma) \otimes H^l(D)} \leq C_{\rho} \| u_{N,p} \|_{L^2_\rho (\Gamma) \otimes H^r(D)}
\]

for \( 1 \leq r \leq \min \{ q + 1, s \} \) whenever \( u_{N,p} \in L^2_\rho (\Gamma) \otimes H^s(D) \) (see [9], Remark 4.4.27).

### 4.2.3 Collocation Error - Full Tensor Method

The evaluation of the last error term \( \| u_N - u_{N,p} \|_{V_p} \) deriving from the approximation in the random space requires a more extensive treatment. We first conduct a one-dimensional analysis using the same notation as in the previous chapters with \( m = 1, \ldots, N \). Let \( \Gamma_m \subset \mathbb{R} \) be a bounded or unbounded domain. Consider the density

\[
\hat{\rho}_m : \Gamma_m \rightarrow \mathbb{R}^+, \quad \hat{\rho}_m(y_m) \leq C(m) e^{-\left( \delta_m y_m \right)^2}
\]

for some \( C_{\max} > 0 \) and \( \delta_m = 0 \) in the case \( \Gamma_m \) is bounded and \( \delta_m > 0 \) when \( \Gamma_m \) is unbounded. Moreover let \( \sigma_m : \Gamma_m \rightarrow \mathbb{R}^+ \) such that

\[
\sigma_m(y_m) \geq C_m e^{-\left( \delta_m y_m \right)^2}
\]

for some \( C_m > 0 \). Observe that this requirement is fulfilled both by a Gaussian weight \( \sigma_m(y_m) = e^{-\left( \mu_m y_m \right)^2} \) for \( \mu_m \leq \delta_m / 2 \) and an exponential weight \( \sigma_m(y_m) = e^{-\alpha_m |y_m|} \) for some \( \alpha_m > 0 \). Recall the functional space

\[
C^0_{\sigma_m}(\Gamma_m; V) = \left\{ v : \Gamma_m \rightarrow V, v \text{ continuous in } y_m, \max_{y_m \in \Gamma_m} \sigma_m(y_m) \| v(y_m) \|_V < \infty \right\}
\]

where \( V \) is an Hilbert space. Under the previous assumptions it easily follows that the following continuous embedding holds

\[
C^0_{\sigma_m}(\Gamma_m; V) \hookrightarrow L^2_\rho (\Gamma_m; V)
\]

and we denote the continuity constant with \( C_1 \).
Denote with $\{y_{m,k}\}_{k=1}^{p_m+1}$ the zeros of the non-trivial polynomial $q_{p_m+1} \in P_{p_m+1}(\Gamma_m)$ orthogonal to the space $P_{p_m}(\Gamma_m)$ with respect to $\hat{\rho}_m$. Denote by $I_{p_m}$ the Lagrange interpolation operator

$$I_{p_m} : C_0(\sigma_m; V) \rightarrow L^2(\Gamma_m; V), \quad I_{p_m} v(y_m) = \sum_{k=1}^{p_m+1} v(y_{m,k}) \ell_{m,k}(y_m)$$

(4.2.4)

where $\ell_{m,k} \in P_{p_m}(\Gamma_m)$, $\ell_{m,k}(y_m) = \prod_{i=1 \atop i \neq k}^{p_m+1} \frac{y_m - y_{m,i}}{y_{m,k} - y_{m,i}}$.

Observe that $\ell_{m,j}(y_{m,i}) = \delta_{ji}$ for $j, i = 1, \ldots, p_m + 1$. In the following we exploit several properties of the above polynomials. The first one is their mutual orthogonality with respect to $\hat{\rho}_m$. Indeed by noting that $\ell_{m,j}(y_{m,i}) = 0$ for all $j, i = 1, \ldots, p_m + 1$ and $i \neq j$, we can conclude that there exists a polynomial $h \in P_{p_m}(\Gamma_m)$ such that

$$\ell_{m,j}(y_m)\ell_{m,i}(y_m) = h(y_m)q_{p_m+1}(y_m).$$

Therefore by assumption on $q_{p_m+1}$ we obtain

$$\int_{\Gamma_m} \ell_{m,j}(y_m)\ell_{m,i}(y_m)\hat{\rho}_m(y_m)dy_m = \int_{\Gamma_m} h(y_m)q_{p_m+1}(y_m)\hat{\rho}_m(y_m)dy_m = 0.$$

The second useful property reads

$$\sum_{k=1}^{p_m+1} \ell_{m,k}(y_m) = 1.$$  

(4.2.5)

This easily follows after observing that $\sum_{k=1}^{p_m+1} \ell_{m,k}(y_m) - 1$ is a polynomial of degree $p_m$ with the $p_m + 1$ roots $y_{m,k}$. Consequently it has to be $\sum_{k=1}^{p_m+1} \ell_{m,k}(y_m) - 1 \equiv 0$.

Recall from Section 3.1 the Gauss quadrature weights

$$\omega_{m,k} = \int_{\Gamma_m} \ell_{m,k}^2(y_m)\hat{\rho}_m(y_m)dy_m.$$

Observe that $\sum_{k=1}^{p_m+1} \omega_{m,k} = 1$. Indeed by orthogonality and equation (4.2.5) we have

$$\sum_{k=1}^{p_m+1} \omega_{m,k} = \int_{\Gamma_m} \left( \sum_{k=1}^{p_m+1} \ell_{m,k}^2(y_m)\hat{\rho}_m(y_m)dy_m \right) \hat{\rho}_m(y_m)dy_m = \int_{\Gamma_m} \left( \sum_{k=1}^{p_m+1} \ell_{m,k}(y_m) \right)^2 \hat{\rho}_m(y_m)dy_m = \int_{\Gamma_m} \hat{\rho}_m(y_m)dy_m = 1.$$
Lemma 4.2.1. The operator \( \mathcal{I}_{\rho_m} : C^0_{\sigma_m}(\Gamma_m; V) \to L^2_{\rho_m}(\Gamma_m; V) \) defined in (4.2.4) is continuous with
\[
\| \mathcal{I}_{\rho_m} v \|_{L^2_{\rho_m}(\Gamma_m; V)} \leq C_2 \| v \|_{C^0_{\sigma_m}(\Gamma_m; V)}.
\]

Proof. For any \( v \in C^0_{\sigma_m}(\Gamma_m; V) \) we have
\[
\| \mathcal{I}_{\rho_m} v \|_{L^2_{\rho_m}(\Gamma_m; V)}^2 = \int_{\Gamma_m} \| \mathcal{I}_{\rho_m} v(y_m) \|^2 \, \hat{\rho}_m(y_m) \, dy_m
\]
\[
= \int_{\Gamma_m} \left( \sum_{k_m = 1}^{p_m+1} v(y_m, k_m) \ell_{m,k_m}(y_m) \right)^2 \, \hat{\rho}_m(y_m) \, dy_m
\]
\[
\leq \int_{\Gamma_m} \sum_{k_m = 1}^{p_m+1} \ell_{m,k_m}(y_m) \ell_{m,k_m}(y_m) \| v(y_m, k_m) \|_V \| v(y_m, k_m) \|_V \, \hat{\rho}_m(y_m) \, dy_m
\]
\[
= \sum_{k_m = 1}^{p_m+1} \| v(y_m, k_m) \|^2 \, \omega_{m,k_m}
\]
\[
\leq \max_{k_m \in \{1, \ldots, p_m+1\}} \sigma_m^2(y_m, k_m) \| v(y_m, k_m) \|^2 \, \sum_{k_m = 1}^{p_m+1} \frac{\omega_{m,k_m}}{\sigma_m^2(y_m, k_m)}
\]
\[
\leq \| v \|^2_{C^0_{\sigma_m}(\Gamma_m; V)} \sum_{k_m = 1}^{p_m+1} \frac{\omega_{m,k_m}}{\sigma_m^2(y_m, k_m)}
\]

(4.2.6)

where we have used the orthogonality property of the Lagrange polynomials. Now we have to possible cases. If \( \Gamma_m \) is bounded then \( \delta_m = 0 \) and consequently \( \sigma_m \geq C_m \). Thus (4.2.6) gives
\[
\| \mathcal{I}_{\rho_m} v \|^2_{L^2_{\rho_m}(\Gamma_m; V)} \leq \frac{1}{C_m^2} \| v \|^2_{C^0_{\sigma_m}(\Gamma_m; V)} \sum_{k_m = 1}^{p_m+1} \omega_{k_m} = \frac{1}{C_m^2} \| v \|^2_{C^0_{\sigma_m}(\Gamma_m; V)}.
\]

In the case of \( \Gamma_m \) unbounded we apply a result from [5, Section 7]. Namely as all the even moments \( \int_{\Gamma_m} y_m^{2n} \hat{\rho}_m(y_m) \, dy_m \) are bounded by the moments of the Gaussian density \( e^{-((\delta_m y_m)^2)} \) by assumption on \( \hat{\rho}_m \), we can conclude
\[
\sum_{k_m = 1}^{p_m+1} \frac{\omega_{m,k_m}}{\sigma_m^2(y_m, k_m)} \xrightarrow{p_m \to \infty} \int_{\Gamma_m} \frac{\hat{\rho}_m(y_m)}{\sigma_m^2(y_m)} \, dy_m \leq \frac{C^{(m)}_m}{C_m^2} \frac{\sqrt{2\pi}}{\delta_m}.
\]

Therefore from (4.2.6)
\[
\| \mathcal{I}_{\rho_m} v \|^2_{L^2_{\rho_m}(\Gamma_m; V)} \leq \| v \|^2_{C^0_{\sigma_m}(\Gamma_m; V)} \sum_{k_m = 1}^{\infty} \frac{\omega_{m,k_m}}{\sigma_m^2(y_m, k_m)} \leq \frac{C^{(m)}_m \sqrt{2\pi}}{C_m^2 \delta_m} \| v \|^2_{C^0_{\sigma_m}(\Gamma_m; V)}.
\]

39
Hence the lemma is proved with continuity constant independent of $p_m$ defined as

$$C_2^2 := \begin{cases} 
\frac{1}{C_m^{(2)}}, & \Gamma_m \text{ is bounded} \\
\frac{C_m^{(2)}\sqrt{2\pi}}{\delta_m}, & \Gamma_m \text{ is unbounded} 
\end{cases}$$

\[ \square \]

**Lemma 4.2.2.** For any function $v \in C^0_{\sigma_m}(\Gamma_m; V)$ the interpolation error satisfies

$$\|v - I_{p_m}v\|_{L^2_{p_m}(\Gamma_m; V)} \leq C_3 \inf_{w \in \mathcal{P}_{p_m}(\Gamma_m) \otimes V} \|v - w\|_{C^0_{\sigma_m}(\Gamma_m; V)}$$

(4.2.7)

where $C_3$ is independent of $p_m$.

**Proof.** Note that for all $w \in \mathcal{P}_{p_m}(\Gamma_m) \otimes V$ it holds that $I_{p_m}w = w$. Thus

$$\|v - I_{p_m}v\|_{L^2_{p_m}(\Gamma_m; V)} \leq \|v - w\|_{L^2_{p_m}(\Gamma_m; V)} + \|I_{p_m}w - I_{p_m}v\|_{L^2_{p_m}(\Gamma_m; V)}$$

$$\leq \|v - w\|_{L^2_{p_m}(\Gamma_m; V)} + \|I_{p_m}(w - v)\|_{L^2_{p_m}(\Gamma_m; V)}$$

$$\leq C_1\|v - w\|_{C^0_{\sigma_m}(\Gamma_m; V)} + C_2\|v - w\|_{C^0_{\sigma_m}(\Gamma_m; V)}$$

$$\leq C_3\|v - w\|_{C^0_{\sigma_m}(\Gamma_m; V)}$$

where in the second-to-last inequality we have used the embedding (4.2.3) and Lemma 4.2.1. The constant $C_3$ is equal to $\max\{C_1, C_2\}$. Since $w \in \mathcal{P}_{p_m}(\Gamma_m) \otimes V$ was arbitrary, the claim follows. \[ \square \]

Now we examine the error on the right hand side of (4.2.7). In particular we consider the case of a function $v : \Gamma_m \to V$ which admits an analytic extension, again denoted with $v$, in the region $\Sigma(\Gamma_m; \tau_m) := \{z \in \mathbb{C} : \text{dist}(z, \Gamma_m) \leq \tau_m\} \subset \mathbb{C}$ for some $\tau_m > 0$. We split the analysis into the two cases of $\Gamma_m$ bounded or unbounded.

Let us first focus on the bounded case where $\sigma_m = 1$.

**Lemma 4.2.3.** Let $v : \Gamma_m \to V$ be a function which admits an analytic extension $v$ in the region $\Sigma(\Gamma_m; \tau_m)$ for some $\tau_m > 0$. Then

$$\inf_{w \in \mathcal{P}_{p_m}(\Gamma_m) \otimes V} \|v - w\|_{C^0(\Gamma_m; V)} \leq \frac{2}{q_m - 1} q_m^{-p_m} \max_{z \in \Sigma(\Gamma_m; \tau_m)} \|v(z)\|_V$$

where $1 < q_m = \frac{2\tau_m}{|\Gamma_m|} + \sqrt{\frac{4\tau_m^2}{|\Gamma_m|^2} + 1}$.

**Proof.** Consider the following change of variable: $g(t) = y_0 + \frac{|\Gamma_m|}{2}t$ with $y_0$ the midpoint of $\Gamma_m$. Consequently $g([-1, 1]) = \Gamma_m$. Let $\tilde{v}(t) := v(g(t))$. Thus $\tilde{v}$ admits an analytic extension in the region

$$\Sigma\left([-1, 1]; \frac{2\tau_m}{|\Gamma_m|}\right).$$

40
The function \( x \mapsto \tilde{v}(\cos x) \) is even, \( 2\pi \)-periodic with continuous derivative. Thus it has a convergent Fourier series

\[
\tilde{v}(\cos x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(kx)
\]

where \( a_k \in V \) is defined as

\[
a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} \tilde{v}(\cos x) \cos(kx) dx.
\]

Define the Chebyshev polynomials \( C_k(\cos x) := \cos(kx) \). Let \( t = \cos x \). We can reformulate the previous expansion in terms of the \( C_k \)'s on \([-1, 1]\) and we get for \( \tilde{v} : [-1, 1] \to V \)

\[
\tilde{v}(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k C_k(t). \tag{4.2.8}
\]

Thanks to [6, Theorem 7] and [7, proof of Theorem 8.1] we know that for all \( \varrho_m > 1 \) such that \( \tilde{v} \) is analytic on \( D_{\varrho_m} := \{ z = y + iw : \frac{y^2}{a^2} + \frac{w^2}{b^2} \leq 1, a = \frac{\varrho_m + \varrho_m^{-1}}{2}, b = \frac{\varrho_m - \varrho_m^{-1}}{2} \} \), it holds that the series (4.2.8) converges within \( D_{\varrho_m} \) and for all \( k \in \mathbb{N} \) the coefficients satisfy the following bound

\[
\|a_k\|_V \leq 2\varrho_m^{-k} \max_{z \in D_{\varrho_m}} \|\tilde{v}(z)\|_V. \tag{4.2.9}
\]

Denote by \( \Pi_{p_m} \) the truncation of Chebyshev expansion (4.2.8) up to \( p_m \). Then

\[
\inf_{w \in P_{p_m}(\Gamma_m) \otimes V} \|v - w\|_{C^0(\Gamma_m; V)} = \inf_{\tilde{w} \in P_{p_m}([-1, 1]) \otimes V} \|\tilde{v} - \tilde{w}\|_{C^0([-1, 1]; V)} \\
\leq \|\tilde{v} - \Pi_{p_m} \tilde{v}\|_{C^0([-1, 1]; V)} \\
= \left\| \sum_{k=p_m+1}^{\infty} a_k C_k(t) \right\|_{C^0([-1, 1]; V)} \\
\leq \sum_{k=p_m+1}^{\infty} \|a_k\|_V \\
\leq 2 \max_{z \in D_{\varrho_m}} \|\tilde{v}(z)\|_V \sum_{k=p_m+1}^{\infty} \varrho_m^{-k} \\
= 2 \max_{z \in D_{\varrho_m}} \|\tilde{v}(z)\|_V \frac{\varrho_m^{-p_m}}{\varrho_m - 1} \\
\leq 2 \max_{z \in \Sigma([-1, 1]; \varrho_m)} \|\tilde{v}(z)\|_V \frac{\varrho_m^{-p_m}}{\varrho_m - 1} \\
= 2 \max_{z \in \Sigma(\Gamma_m; \varrho_m)} \|v(z)\|_V \frac{\varrho_m^{-p_m}}{\varrho_m - 1}
\]
where we have used in the second inequality the fact that $\max_{t \in [-1,1]} |C_k(t)| = 1$ and in the third inequality the bound (4.2.9).

Finally observe that the largest ellipse that can be drawn inside $\Sigma \left([0,1]; \frac{2\tau m}{|\Gamma_m|}\right)$ is obtained by equating the minor semi-axis of $\partial D_{\varrho m}$, the boundary of $D_{\varrho m}$, with the radius $\frac{2\tau m}{|\Gamma_m|}$, i.e.

$$\frac{\varrho_m - \varrho_m^{-1}}{2} = \frac{2\tau m}{|\Gamma_m|}.$$ 

Solving the equation we find

$$\varrho_m = \frac{2\tau m}{|\Gamma_m|} + \sqrt{\frac{4\tau_m^2}{|\Gamma_m|^2} + 1}.$$ 

Now we turn to the case of $\Gamma_m$ unbounded, i.e. $\Gamma_m = \mathbb{R}$. We first present a result concerning Hermite polynomials. Let $H_n(y) \in \mathcal{P}_n(\mathbb{R})$ denote the normalized Hermite polynomials

$$H_n(y) = \frac{1}{\sqrt{\sqrt{\pi}2^n n!}} (-1)^n e^{y^2} \partial^n \partial y^n (e^{-y^2})$$

and $h_n$ the Hermite functions

$$h_n(y) = e^{-\frac{y^2}{2}} H_n(y).$$

From [13] we know that Hermite polynomials form a complete orthonormal basis of $L^2(\mathbb{R})$ with respect to the weight $e^{-y^2}$, i.e.

$$\int_{\mathbb{R}} H_n(y) H_m(y) e^{-y^2} dy = \delta_{nm}.$$ 

It has been proved in [8, Theorem 1] that the following holds true.

**Lemma 4.2.4.** Let $f$ be an analytic function in $S(\mathbb{R}; \tau) := \{ z = y + iw : w \in [-\tau, \tau] \} \subset \mathbb{C}$. The Hermite-Fourier series

$$\sum_{n=0}^{\infty} f_n h_n(z), \quad f_n = \int_{\mathbb{R}} f(y) h_n(y) dy$$

exists and converges to $f$ in $S(\mathbb{R}; \tau)$ if and only if for every $\beta \in [0, \tau)$ there exists a finite positive constant $C(\beta)$ such that

$$|f(y + iw)| \leq C(\beta) e^{-|y|\sqrt{\beta^2 - w^2}}, \quad y \in \mathbb{R}, \quad w \in [-\beta, \beta].$$

Moreover it holds that

$$|f_n| \leq C e^{-\tau \sqrt{2n+1}}.$$ (4.2.11)
Recall from the beginning of the section the two weights
\[ G_m(y) := e^{-\left(\frac{\|y_m\|^2}{4}\right)}, \quad \sigma_m(y) = e^{-\alpha_m|y_m|} \]
for some \(\alpha_m, \delta_m > 0\). Clearly \(C^0_{\sigma_m}(\Gamma_m; V) \hookrightarrow C^0_{G_m}(\Gamma_m; V)\).

**Lemma 4.2.5.** Let \(v \in C^0_{\sigma_m}(\Gamma_m; V)\) which admits an analytic extension in \(\Sigma(\mathbb{R}; \tau_m) \subset \mathbb{C}\) for some \(\tau_m > 0\) and such that for all \(z = y + iw \in \Sigma(\mathbb{R}; \tau_m)\) it holds
\[ \sigma_m(y)\|v(z)\|_V \leq C_v(\tau_m) \]
for some positive finite constant \(C_v(\tau_m)\). Then, for any \(\delta_m > 0\), there exist a constant \(K\), independent of \(p_m\), and a function \(\chi(p_m) = O(\sqrt{p_m})\) such that
\[ \inf_{w \in \mathcal{T}_{p_m}(\Gamma_m) \cap V} \|v - w\|_{C^0_{\sigma_m}(\Gamma_m; V)} \leq K \Theta(p_m)e^{-\tau_m \delta_m \sqrt{p_m}}. \]

**Proof.** Consider the following change of variable: \(g(t) = \frac{\sqrt{2}}{\delta_m} t\). Let \(\tilde{v}(t) := v(g(t))\). Thus \(\tilde{v}\) admits an analytic extension in the region
\[ \Sigma\left(\mathbb{R}; \frac{\tau_m \delta_m}{\sqrt{2}}\right). \]
Moreover observe that \(\tilde{v} \in C^0_{\sigma_m}(\Gamma_m; V)\) with \(\tilde{\sigma}_m(t) = e^{-\sqrt{2} \frac{\delta_m}{\tau_m} |t|}\). We consider the expansion of \(\tilde{v}\) in Hermite polynomials
\[ \tilde{v}(t) = \sum_{n=0}^{\infty} v_n H_n(t), \quad v_n = \int_{\mathbb{R}} \tilde{v}(t)H_n(t)e^{-t^2} dt. \quad (4.2.12) \]
Now define \(f(z) := \tilde{v}(z)e^{-\frac{z^2}{2}}\). Note that the Hermite-Fourier series of \(f\) defined in (4.2.10) has the same coefficients as the expansion (4.2.12). Indeed
\[ f_n = \int_{\mathbb{R}} f(t)h_n(t) dt = \int_{\mathbb{R}} \tilde{v}(t)e^{-\frac{t^2}{2}} e^{-\frac{t^2}{2}} H_n(t) dt = v_n. \]
As a product of analytic functions, \(f\) is also analytic in \(\Sigma\left(\mathbb{R}; \frac{\tau_m \delta_m}{\sqrt{2}}\right)\). Moreover
\[ \|f(y + iw)\|_V = \left| e^{-\frac{(y+iw)^2}{2}} \right| \|\tilde{v}(z)\|_V \leq e^{-\frac{y^2-w^2}{2}} \frac{1}{\sigma_m(y)} C_v(\tau_m) = e^{-\frac{y^2-w^2}{2}} e^{\sqrt{2} \frac{\delta_m}{\tau_m} |y|} C_v(\tau_m). \]
For \(\beta_m \in \left[0, \frac{\tau_m \delta_m}{\sqrt{2}}\right]\) define
\[ C(\beta_m) := \max_{y \in \mathbb{R}} \max_{w \in [-\beta_m, \beta_m]} \exp \left\{ -\frac{y^2-w^2}{2} + \sqrt{2} \frac{\alpha_m}{\delta_m} |y| + |y| \sqrt{\beta_m^2-w^2} \right\}. \]

43
Note that this is a bounded constant for all $\beta_m$. Observe also that
\[
\|f(y + iw)\|_V \leq C_v(\tau_m) C(\beta_m) e^{-|y|\sqrt{\beta_m^2 - w^2}}.
\]
Therefore $f$ satisfies the assumptions of Lemma 4.2.4. Hence the Hermite-Fourier series of $f$ converges in $\Sigma \left( \mathbb{R}; \frac{\tau_m \delta_m}{\sqrt{2}} \right)$ and the coefficients $f_n$ satisfies the bound (4.2.11). Consequently
\[
\|v_n\|_V \leq C e^{-\frac{\tau_m \delta_m}{\sqrt{2}} \sqrt{2(n+1)}}.
\]
Denote by $\Pi_{p_m}$ the truncation of the Hermite expansion (4.2.12) up to $p_m$. Then
\[
\inf_{w \in \mathcal{P}_{p_m}(\Gamma_m) \otimes V} \|v - w\|_{C^0_{G_m}(\Gamma_m;V)} = \inf_{w \in \mathcal{P}_{p_m}(\Gamma_m) \otimes V} \|\hat{v} - \hat{w}\|_{C^0_{G_m}(\Gamma_m;V)} \leq \max_{t \in \mathbb{R}} \left\| \sum_{n=p_m+1}^{\infty} v_n h_n(t) e^{-t^2/2} \right\|_V.
\]
From [10, bound (2.2)] we know that $|h_n(t)| \leq 1$ for all $t \in \mathbb{R}$ and all $n$. Thus
\[
\inf_{w \in \mathcal{P}_{p_m}(\Gamma_m) \otimes V} \|v - w\|_{C^0_{G_m}(\Gamma_m;V)} \leq \sum_{n=p_m+1}^{\infty} \|v_n\|_V \leq C \sum_{n=p_m+1}^{\infty} e^{-\frac{\tau_m \delta_m}{\sqrt{2}} \sqrt{2(n+1)}}.
\]
Finally we use the following result given in [11, Lemma A.2]: for $r \in \mathbb{R}^+$, $r < 1$ it holds
\[
\sum_{n=p+1}^{\infty} r^{\sqrt{2n+1}} \leq \left( \frac{2\sqrt{p+1}}{r \sqrt{2} (1 - r \sqrt{2})} + \mathcal{O}(1) \right) r^{\sqrt{2} p}.
\]
\[
\text{We can now proceed to prove the following}
\]
\textbf{Theorem 4.2.6.} Under the assumptions of Section 4.1, there exist positive constants $(r_m)_{m=1}^{N}$ and $C$, independent of $p = (p_1, \ldots, p_N)$, such that the approximation error of the full tensor collocation method satisfies
\[
\|u_N - \mathcal{I}^{\text{full}}_N u_N\|_{V_p} \leq C \sum_{m=1}^{N} \chi_m(p_m) e^{-r_m \beta_m^p n_m}.
\]
where

\[
\theta_m = \begin{cases} 
1, & \text{if } \Gamma_m \text{ is bounded} \\
\frac{1}{2}, & \text{if } \Gamma_m \text{ is unbounded}
\end{cases},
\chi_m(p_m) = \begin{cases} 
1, & \text{if } \Gamma_m \text{ is bounded} \\
O(\sqrt{p_m}), & \text{if } \Gamma_m \text{ is unbounded}
\end{cases},
\]

\[
r_m = \begin{cases} 
\log\left(\frac{2\tau_m}{|\Gamma_m|} \right) + \sqrt{1 + \frac{4\tau_m^2}{|\Gamma_m|^2}} , & \text{if } \Gamma_m \text{ is bounded} \\
\tau_m \delta_m , & \text{if } \Gamma_m \text{ is unbounded}
\end{cases},
\]

with \( \tau_m \) defined in Lemma 4.1.2 and \( \delta_m \) as in (4.1.1).

**Proof.** Firstly observe that by passing to the probability space \( L_\rho^2(\Gamma) \) we have

\[
\|u_N - \mathcal{I}^\text{full} u_N\|_{V_\rho} = \|u_N - \mathcal{I}^\text{full} u_N\|_{V_\rho},
\]

where \( V_\rho = L_\rho^2(\Gamma) \otimes H_0^1(D) \). Consequently we want to provide an upper-bound for

\[
\|u_N - \mathcal{I}^\text{full} u_N\|_{V_\rho}.
\]

In order to lighten the notation we will drop the subscript \( N \).

Now observe that from (4.1.2) we have

\[
\|u - \mathcal{I}^\text{full} u\|_{L_\rho^2(\Gamma) \otimes H_0^1(D)} \leq \left\| \rho \right\|^{1/2} \left\| u - \mathcal{I}^\text{full} u\|_{L_\rho^2(\Gamma) \otimes H_0^1(D)} \right\|.
\]

To examine the error in the right-hand side we want to exploit the one-dimensional analysis we have conducted in Section 4.2.3.

The term \( u - \mathcal{I}^\text{full} u \) is an interpolation error. Indeed recall that \( \mathcal{I}^\text{full} = \bigotimes_{m=1}^N \mathcal{I}_{p_m} \) and \( \mathcal{I}_{p_m} \) is the one-dimensional interpolant defined in (3.1.3).

We define the following interpolation operator acting on the last \( N - j + 1 \) directions with \( j = 1, \ldots, N \)

\[
\mathcal{I}_{p_j} v(y, x) = \sum_{k_j=1}^{p_j+1} \cdots \sum_{k_N=1}^{p_N+1} v(y_1, \ldots, y_{j-1}, y_j, k_j, \ldots, y_N, k_N, x) \ell_{j,k_j}(y_j) \cdots \ell_{N,k_N}(y_N).
\]

Equivalently \( \mathcal{I}_{p_j} = \mathcal{I}_{p_1} \otimes \cdots \otimes \mathcal{I}_{p_N} \). Let us consider the first direction. Since

\[
\mathcal{I}^\text{full} = \mathcal{I}_{p_1} \otimes \mathcal{I}_{p_2},
\]

we get

\[
\|u - \mathcal{I}^\text{full} u\|_{L_\rho^2(\Gamma) \otimes H_0^1(D)} = \|u - \mathcal{I}_{p_1} u + \mathcal{I}_{p_1} u - (\mathcal{I}_{p_1} \otimes \mathcal{I}_{p_2}) u\|_{L_\rho^2(\Gamma) \otimes H_0^1(D)} \\
\leq \|u - \mathcal{I}_{p_1} u\|_{L_\rho^2(\Gamma) \otimes H_0^1(D)} + \|\mathcal{I}_{p_1} (u - \mathcal{I}_{p_2} u)\|_{L_\rho^2(\Gamma) \otimes H_0^1(D)}.
\]
We introduce the following notation:

\[ \Gamma_j^* := \prod_{i=j}^{N} \Gamma_i, \quad \hat{\rho}_j^* := \prod_{i=j}^{N} \hat{\rho}_i, \quad \sigma_j^* := \prod_{i=j}^{N} \sigma_i. \]

We view \( u \) as a function \( u : \Gamma_1 \to L_{\hat{\rho}_2}^p(\Gamma_2^*) \otimes H_0^1(D) \). Then \( u \in L_{\hat{\rho}_1}^2(\Gamma_1; L_{\hat{\rho}_2}^2(\Gamma_2^*) \otimes H_0^1(D)) \).

This allows us to write

\[ \inf_{w \in \mathcal{P}_{\hat{\rho}_1}(\Gamma_1) \otimes V} \| u - w \|_{C_{G_1}^0(\Gamma_1; V)} \leq \left\{ \begin{array}{ll}
\frac{2}{\theta_1 - 1} \max_{z \in \Sigma(\Gamma_1; \tau_1)} \| u(z) \|_V e^{-p_1 \log \theta_1}, & \Gamma_1 \text{ is bounded} \\
K_1 O(\sqrt{\tau_1}) e^{-\tau_1 \delta_1 \sqrt{\tau_1}}, & \Gamma_1 \text{ is unbounded}
\end{array} \right. \]

where \( \theta_1 = \frac{2 \tau_1}{\tau_1} + \sqrt{1 + \frac{4 \tau_1^2}{\tau_1}} \) and \( C_1 := \max \left\{ K_1, \frac{2}{\theta_1 - 1} \max_{z \in \Sigma(\Gamma_1; \tau_1)} \| u(z) \|_V \right\} \) is independent of \( p_1 \). Note that in both cases we need the regularity results as stated in Lemma 4.1.2. Moreover by Lemma 4.1.1 the assumption \( u \in C_{G_1}^0(\Gamma_1; V) \) of Lemma 4.2.5 is fulfilled.

It remains to bound the term in (4.2.13). We can do that by using Lemma 4.2.1 and we get

\[ \|u - \mathcal{I}_{\hat{\rho}_1} u\|_{L_{\hat{\rho}_1}^2(\Gamma_1; V)} \leq \sqrt{C_2} \| u - \mathcal{I}_{\hat{\rho}_2} u\|_{C_{G_1}^0(\Gamma_1; V)}. \]

Observe that for the right-hand side we have

\[ \|u - \mathcal{I}_{\hat{\rho}_2} u\|_{C_{G_1}^0(\Gamma_1; L_{\hat{\rho}_2}^2(\Gamma_2^*) \otimes H_0^1(D))} = \max_{y_1 \in \Gamma_1} \|u - \mathcal{I}_{\hat{\rho}_2} u\|_{L_{\hat{\rho}_2}^2(\Gamma_2^*) \otimes H_0^1(D)} \]

\[ \leq \max_{y_1 \in \Gamma_1} \|u - \mathcal{I}_{\hat{\rho}_2} u\|_{L_{\hat{\rho}_2}^2(\Gamma_2^*) \otimes H_0^1(D)}. \]
Therefore we can proceed iteratively considering $I_{p_2}$ and $I_{p_3}$ such that $I_{p_2} = I_{p_2} \otimes I_{p_3}$ and (4.2.15) can be bounded, uniformly in $\Gamma_1$, in the following way

$$\|u - I_{p_2}u\|_{L^2(\Gamma_1) \otimes H^1_0(D)} \leq \|u - I_{p_2}u\|_{L^2(\Gamma_2; L^2_\rho_3(\Gamma_3) \otimes H^1_0(D))} + \|I_{p_2}(u - I_{p_2}u)\|_{L^2(\Gamma_2; L^2_\rho_3(\Gamma_3) \otimes H^1_0(D))}.$$ 

The most-right-hand side in the previous display will be again bounded uniformly in $\Gamma_2$ and so on considering all the remaining directions $y_3, \ldots, y_N$. At each step for $j = 2, \ldots, N - 1$ we define

$$C_j = \max \left\{ K_j, \frac{2}{\rho_j - 1} \max_{\Sigma(\Gamma_j; \tau_j)} \|u\|_{L^2(\Gamma_{j+1}; H^1_0(D))} \right\}$$

and

$$C_N = \max \left\{ K_N, \frac{2}{\rho_N - 1} \max_{\Sigma(\Gamma; \tau)} \|u\|_{H^1_0(D)} \right\}.$$ 

The constant $C$ appearing in the statement will be then

$$C = \max \left\{ C_1, \max_{y_1 \in \Gamma_1} C_2, \ldots, \max_{y_1 \in \Gamma_1} \cdots \max_{y_{N-1} \in \Gamma_{N-1}} C_N \right\}.$$ 

\[ \square \]

### 4.2.4 Convergence Result

We can collect all the results obtained so far in the following

**Theorem 4.2.7.** If the following assumptions are satisfied:

(A1) $a \in L^2_\rho(\Omega) \otimes L^2(D)$.

(A2) $a$ is uniformly bounded from below by $a_{\min} > 0$.

(A3) $f \in L^2_\rho(\Omega) \otimes L^2(D)$.

(A4) $f \in C^0_\sigma(\Gamma; L^2(D))$.

(A5) $\forall y \in \Gamma, \rho(y) \leq C_\rho e^{-\sum_{m=1}^N (\delta_m y_m)^2}$ for some $C_\rho > 0$.

(A6) $\forall y \in \Gamma, \exists y_m < \infty : \left\| \frac{\partial^k y_m a(y)}{a(y)} \right\|_{L^\infty(D)} \leq \gamma^k m!$ and $\left\| \frac{\partial^k y_m f(y)}{1 + f(y)} \right\|_{L^2(D)} \leq \gamma^k m!$.

(A7) $(Y_m)_m$ is uniformly bounded in $L^\infty(\Omega)$.

(A8) $u_{N,p} \in L^2_\rho(\Gamma) \otimes (H^s(D) \cap H^1_0(D))$ for $s \geq 1$.

(A9) $\forall y_m \in \Gamma_m, \sigma_m(y_m) \geq C_m e^{-\frac{(\delta_m y_m)^2}{4}}$ for some $C_m > 0$. 

47
Then for $1 \leq r \leq \min\{q + 1, s\}$ and $q$ the degree of the polynomial approximation in $D$, the following error bound for the stochastic full tensor collocation approximation holds true

$$
\|u - u_{N,p,h}\|_{L^2_p(\Omega) \otimes H^1(D)} \lesssim \sum_{m \geq N+1} \mu_m + \|f\|_{L^2_p(\Omega) \otimes L^2(D)} \sum_{m \geq N+1} \sqrt{\lambda_m} \|\phi_m\|_{L^\infty(D)}
+ h^{r-1}\|u_{N,p}\|_{L^2(\Gamma) \otimes H^r(D)}
+ \sum_{m=1}^N \chi_m(p_m)e^{-r_m p_m^2} \tag{4.2.16}
$$

where $\mu_m$, $\lambda_m$, and $\phi_m$ are as in Section 4.2.1, $h$ is the mesh size introduced in Section 4.2.2 and $\chi_m$, $r_m$ and $\theta_m$ are defined in Theorem 4.2.6.

**Proof.** The first two terms on the right-hand side of (4.2.16) are related to the truncation error and the bound can be found in Section 4.2.1. The second line of the claim concerns the spatial approximation in $D$ and it has been examined in Section 4.2.2. The last term controls the convergence in the random domain $\Omega$ and the result is given in Theorem 4.2.6.

### 4.2.5 Collocation Error - Sparse Tensor Method

We have an analogue result of Theorem 4.2.6 when we adopt a sparse collocation grid instead of the full one. For a more detailed proof see [12]. We adopt the same notation as in Section 3.2.

**Theorem 4.2.8.** Under the assumptions of Section 4.1, in the case of $\Gamma$ bounded, the approximation error of the isotropic Smolyak sparse tensor collocation method satisfies

$$
\|u_N - S_{t,g}^w u_N\|_{V_p} \leq C(r_{\min}, N)w e^{-\frac{r_{\min}}{2}w}
$$

where $r_{\min} := \min_{1 \leq m \leq N} \min_{y_m \in \Gamma_m} r_m$, $r_m$ is defined in Theorem 4.2.6 and the constant $C(r_{\min}, N)$ does not depend on $w$.

**Proof.** As in the proof of Theorem 4.2.6, we have

$$
\|u_N - S_{t,g}^w u_N\|_{V_p} = \|u_N - S_{t,g}^w u_N\|_{V_p} \leq \left\|\frac{\rho}{\hat{\rho}}\right\|_{L^\infty(\Gamma)}^{1/2} \|u_N - S_{t,g}^w u_N\|_{V_p}.
$$

For convenience we drop the subscript $N$. The proof is based on a one-dimensional argument.

Let $w \in \mathbb{N}$ and recall the operator (3.2.2)

$$
S_{t,g}^w = \sum_{i \in \mathbb{N}^N} \bigotimes_{m=1}^N \Delta^i_{m(w)}
$$

48
where
\[ t(i) = \begin{cases} 1, & i = 1 \\ 2^{i-1} + 1, & i \geq 2 \end{cases}, \quad g(i) = \sum_{m=1}^{N} (i_m - 1). \]

Let \( m \in \{1, \ldots, N\} \) and \( V = L^2_{\rho_m}(\Gamma_m^*) \otimes H^1(D) \) with the notation as in Section 4.1. From Lemma 4.2.2 we have for the one-dimensional operator (3.2.1)
\[
\| u - I_t(1_m) u \|_{L^2_{\rho_m}(\Gamma_m;V)} \leq C_3 \inf_{w \in \mathcal{P}_{t(1_m)-1}(\Gamma_m) \otimes V} \| u - w \|_{C^0(\sigma_m;\Gamma_m;V)}.
\]

From Lemma 4.2.3 we also have
\[
\inf_{w \in \mathcal{P}_{t(1_m)-1}(\Gamma_m) \otimes V} \| u - w \|_{C^0(\sigma_m;\Gamma_m;V)} \leq 2 e^{r_{1_m} - 1} M_m(u) e^{-[t(1_m)-1]r_m}
\]

where
\[
M_m(u) := \max_{z \in \Sigma(\Gamma_m;\tau_m)} \| u(z) \|_V, \quad r_m = \log \left( \frac{2 \tau_m}{|\Gamma_m|} + \sqrt{1 + \frac{4 \tau_m^2}{|\Gamma_m|^2}} \right)
\]

and \( \tau_m \) is defined in Lemma (4.1.2).

Now by definition of the difference operator (3.2) we get
\[
\| \Delta_m^{t(1_m)} u \|_{L^2_{\rho_m}(\Gamma_m;V)} = \| (I_{t(1_m)} - I_{t(1_m)-1}) u \|_{L^2_{\rho_m}(\Gamma_m;V)} \leq \| u - I_{t(1_m)} u \|_{L^2_{\rho_m}(\Gamma_m;V)} + \| u - I_{t(1_m)-1} u \|_{L^2_{\rho_m}(\Gamma_m;V)} \leq 2 e^{r_{1_m} - 1} M_m(u) e^{-[t(1_m)-1]r_m}.
\]

Let \( \text{Id} : \Gamma \to \Gamma \) the identity operator on \( \Gamma \). Observe that \( u = \sum_{i \in \mathcal{N}} \bigotimes_{m=1}^{N} \Delta_m^{t(1_m)} u \).

Hence
\[
\| (\text{Id} - \sum_{|i| \leq w} \bigotimes_{m=1}^{N} \Delta_m^{t(1_m)} u) \|_{L^2_{\rho_m}(\Gamma_m;H^1(D))} \leq \sum_{|i| > w} \prod_{m=1}^{N} \| \Delta_m^{t(1_m)} u \|_{L^2_{\rho_m}(\Gamma_m;V)} \leq \sum_{|i| > w} \prod_{m=1}^{N} \frac{4}{e^{r_{1_m} - 1} M_m(u)} e^{-[t(1_m)-1]r_m}.
\]

Define \( M(u) := \max_{1 \leq m \leq N} \max_{y_m \in \Gamma_m} M_m(u) \text{ and } r_{\min} = \min_{1 \leq m \leq N} \min_{y_m \in \Gamma_m} r_m. \)
Then for $\tilde{C} := \frac{4}{e r_{\min} - 1}$ we obtain

\[
\|u - \sum_{|i| \leq w} \bigotimes_{m=1}^{N} \Delta_m^{l(i_m+1)} u\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq \tilde{C}^N M(u)^N \sum_{|i| > w} e^{-r_{\min} \sum_{m=1}^{N} l(i_m) - 1} \\
= \tilde{C}^N M(u)^N \sum_{|i| > w} e^{r_{\min} \sum_{m=1}^{N} 2^{i_m} - 1} \\
= \tilde{C}^N M(u)^N \sum_{|i| > w} e^{r_{\min} \sum_{m=1}^{N} 2^{i_m}}.
\]

From the fact that for any $i \in \mathbb{N}_+$ the inequality $2^i > i$ holds, we get

\[
\|u - \sum_{|i| \leq w} \bigotimes_{m=1}^{N} \Delta_m^{l(i_m+1)} u\|_{L^2_\rho(\Gamma; H^1_0(D))} \leq C(r_{\min}, N) \sum_{|i| > w} e^{r_{\min} \sum_{m=1}^{N} 2^{i_m}} \\
\leq C(r_{\min}, N) \sum_{|i| > w} e^{r_{\min} \sum_{m=1}^{N} i_m} \\
= C(r_{\min}, N) \sum_{|i| > w} e^{r_{\min} |i|} \\
\simeq C(r_{\min}, N) w e^{-r_{\min} w}.
\]

where $C(r_{\min}, N) = [\tilde{C} M(u)]^N = \left[ \frac{4M(u)}{e^{r_{\min}} - 1} \right]^N$. \hfill \qed

**Remark 4.** A related result in the case of $\Gamma$ unbounded holds. In particular in the proof Lemma 4.2.5 has to be invoked instead of Lemma 4.2.3.

Now we want to relate the number of collocation points in the Smolyak sparse grid to the level $w$.

**Lemma 4.2.9.** The cardinality of the Smolyak sparse tensor grid at level $w$ satisfies the following upper bound

\[
\eta = \eta(w, N) \preceq 2^w w^{N-1}
\]

where the omitted constant is of the order $2^N$. 

50
Proof. By definition of Smolyak operator we have

$$\eta \leq \sum_{i \in \mathbb{N}^N : |i| \leq w} \prod_{m=1}^{N} t(i_m + 1) = \sum_{j=0}^{w} \sum_{|i| = j}^{N} \prod_{m=1}^{N} t(i_m + 1) = \sum_{j=0}^{w} \sum_{|i| = j}^{N} (2^{i_m} + 1) \approx \sum_{j=0}^{w} 2^{|i|} = \sum_{j=0}^{w} 2^j \# \{ i \in \mathbb{N}^N_+ : |i| = j \}

= \sum_{j=0}^{w} 2^j \left( \frac{j+N-1}{N-1} \right).$$

Observe that \( \left( \frac{j+N-1}{N-1} \right) \leq \frac{(j+N-1)^{N-1}}{(N-1)!} \approx j^{N-1}. \) On the other hand \( \left( \frac{j+N-1}{N-1} \right) \geq \frac{j^{N-1}}{(N-1)!} \approx j^{N-1}. \) Therefore we get

$$\eta \leq \sum_{j=0}^{w} 2^j \cdot j^{N-1} \approx 2^w \cdot w^{N-1}.$$ 

\( \square \)

### 4.3 Convergence of Moments

The main goal of the implementations which are material of the next chapter is the computation of the mean value and possibly higher moments of the final approximation \( u_{N,p,h} \). Therefore we are interested in providing bounds for the error in the first two moments with respect to some suitable norm.

**Lemma 4.3.1.** For \( V(D) := L^2(D) \) or \( V(D) := H^1_0(D) \) it holds

$$\| E(u - u_{N,p,h}) \|_{V(D)} \leq \| u - u_{N,p,h} \|_{L^2_0(\Omega) \otimes V(D)}.$$

**Proof.** Let us write \( \varepsilon := u - u_{N,p,h} \). We start with the case \( V(D) = L^2(D) \). By Jensen’s inequality we have

$$\| E(\varepsilon) \|_{L^2(D)} \leq \left( \int_D E(\varepsilon^2) \right)^{1/2} = \| \varepsilon \|_{L^2_0(\Omega) \otimes L^2(D)}$$

and we have the claim for \( L^2(D) \). Now consider the case \( V(D) = H^1_0(D) \). Again by Jensen’s inequality we get

$$\| E(\varepsilon) \|_{H^1_0(D)} = \left( \int_D |\nabla E(\varepsilon)|^2 \right)^{1/2} \leq \left( \int_D E(|\nabla \varepsilon|^2) \right)^{1/2} = \| \varepsilon \|_{L^2_0(\Omega) \otimes H^1_0(D)}.$$

\( \square \)

For a proof of the following lemma we refer the reader to [11, Lemma 4.8].
Lemma 4.3.2.
\[
\|E(u^2 - u_{N,p,h}^2)\|_{L^1(D)} \leq C\|u - u_{N,p,h}\|_{L^2_p(\Omega) \otimes L^2(D)}
\]
with constant \(C\) independent of \(h\) and \(p\).

Remark 5. In order to estimate the convergence rate of the error of higher moments or the convergence rate of the error of the second moment in stronger norms, we need more regularity assumptions on the solution.

4.4 Anisotropic Sparse Method - Selection of Weights \(\alpha\)

Recall from the Section 4.2.3 the two Lemmas 4.2.3 and 4.2.5 which state that in a univariate case, assuming analyticity of the solution in the random space, the error of the approximation by polynomials of the dependence on each random variable decays exponentially fast in the polynomial degree. The size \(\tau_m\) of the analyticity region depends in general on the direction \(y_m\) (cf. Lemma 4.1.2). Therefore the decay coefficients in both the lemmas will depend on the same direction as well, respectively for the bounded and unbounded case

\[
 r_m = \log \left( \frac{2\tau_m}{|G_m|} + \sqrt{\frac{4\tau_m^2}{|G_m|^2} + 1} \right), \quad r_m = \tau_m\delta_m.
\]

As we have already remarked in Section 3.2.1, the main point of the anisotropic version of the sparse collocation method is to increase the number of points in those directions where the convergence rate is poor, i.e. \(r_m\) small. One possible choice to relate the weights \(\alpha = (\alpha_1, ..., \alpha_N)\) to the rate \(r = (r_1, ..., r_N)\) is the following

\[
\alpha_m = r_m, \quad \forall m \in \{1, ..., N\}.
\]
5 Numerical Implementation

The last part of this work is devoted to the implementation of the stochastic collocation method in its isotropic version with the aim to validate the theoretical results we have proved in the previous chapters.

5.1 Example

This example is taken from [12]. We analyse the following elliptic boundary value problem
\[
\begin{align*}
-\nabla \cdot (a(\omega, x) \nabla u(\omega, x)) &= f(\omega, x), \quad x \in D \\
u(\omega, x) &= 0, \quad x \in \partial D
\end{align*}
\]
where \(D = [0, 1]^2 \subset \mathbb{R}^2\) and \(x = (x_1, x_2)\). The coefficient \(f\) is deterministic, i.e. \(f(x) = \cos(x_1) \sin(x_2)\).

On the other hand the truncated diffusion coefficient \(a_N\) is given by
\[
\log(a_N(\omega, x) - 0.5) = 1 + Y_1(\omega) \left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} + \sum_{m=2}^{N} \lambda_m \phi_m(x) Y_m(\omega) \tag{5.1.1}
\]
where for \(m \geq 2\)
\[
\lambda_m = (\sqrt{\pi}L)^{1/2} \exp \left( -\frac{\left( \left( \frac{m}{2} \right)^2 \pi L \right)^2}{8} \right), \quad \phi_m(x) = \begin{cases} 
\sin \left( \frac{\pi x_1}{L} \right), & m \text{ even} \\
\cos \left( \frac{\pi x_1}{L} \right), & m \text{ odd}
\end{cases}
\]

It can be shown (see [12]) that (5.1.1) is the truncation of a random field with covariance
\[
V_{\log(a-0.5)}(x, x') = \exp \left( -\frac{(x_1 - x'_1)^2}{L_c^2} \right).
\]
Here \(L_c\) denotes a physical correlation length. This means that the random fields \(\log[a(x) - 0.5]\) and \(\log[a(x') - 0.5]\) are only very weakly correlated for \(|x - x'| \gg L_c\). So we take \(L\) and \(L_p\) in the previous displays as \(L_p = \max\{1, 2L_c\}\) and \(L = L_c/L_p\). In our computations we will fix \(L_c = \frac{1}{32}\).

We consider independent random variables \((Y_m)_{m=1}^N\) which are uniformly distributed on the interval \([-\sqrt{3}, \sqrt{3}]\). Thus the stochastic domain \(\Gamma\) is given by \([-\sqrt{3}, \sqrt{3}]^N\). Moreover the collocation points in this case will be the roots of the Legendre polynomials.
The spatial approximation is conducted applying a finite element method which employs piecewise linear polynomials over a uniform triangulation of $D$ of mesh size $\sqrt{2}/n$, $n \in \mathbb{N}$ (see Figure 5.1.1 for an example).

In this section we are particularly interested in studying the convergence of the collocation error as described in Section 4.2.3 and therefore we fix both the dimension $N$ of the truncation and the dimension $n$ of the finite element space. For $N = 2$ and $n = 8$, the first two statistics of the approximated solution are shown in Figure 5.1.2.

They have been computed as described in (3.1.5) with $p = (5, 5)$. Reasonably the same results can be recovered by means of the Smolyak formula (3.2.7).

To investigate the convergence of the full tensor product method we vary the polynomial degree of the space $P_p(\Gamma)$ defined in Section 3.1. We may estimate the error in each direction separately. Namely we choose the $m^{th}$ direction, we keep fixed all
the other ones and we plot the $H^1_0$-norm of the error between $u_{N,p,h}$ and $u_{N,p,h}$ where $\bar{p} = (p_1, ..., p_m + 1, ..., p_N)$. Thus we study the decay of

$$\|\mathbb{E}_{\bar{p}}(u_{N,p,h}) - \mathbb{E}_{\bar{p}}(u_{N,p,h})\|_{H^1_0(D)}$$

versus $p_m$ for $m = 1$ and $m = 2$, respectively.

![Figure 5.1.3: The relative $H^1_0(D)$-norm error of the expectation of $u_{N,p,h}$ when we modify the polynomial degree $p_m$.](image)

The numerical results reflect the theoretical ones. Indeed from Lemma 4.3.1, we expect the error to decrease exponentially fast as the polynomial degree increases.

Now we investigate the computational error of the expectation in the $L^2(D)$-norm when we increase both the polynomial degree and the number of random variables in the Karhunen-Loève truncated expansion (5.1.1). In this example we have chosen equal degree of the polynomial space in all the $N$ directions, i.e. $p_m = p$ for all $m = 1, ..., N$ and at each step we have increased $p$. Figure 5.1.4 exhibits the error versus the number of collocation points $N_p$ which is given by

$$N_p = (p + 1)^N.$$ 

The plot reveals, in accordance with Theorem 4.2.6, (sub)exponential convergence of the error when the polynomial degree increases. On the other hand, the convergence rate deteriorates by a factor of $\frac{1}{N}$ as $N$ increases. Indeed in terms of the number of points $N_p$ the convergence rate behaves like

$$C(r_{\min}, N)e^{-r_{\min}N^{1/N}}.$$ 

Hence the log of the error decays like $-N_p^{1/N}$.

We may conduct a similar analysis for the sparse tensor product method. In this case we will modify the level $w$ which indirectly acts on the polynomial degree as depicted in (3.2.6). Consequently we will have that the total number of collocation points used
Figure 5.1.4: The relative $L^2(D)$-norm error of the expectation of the final approximation computed by the full tensor product rule for $N = 3, 5, 6$.

at each step is equal to

$$N_p = \sum_{i \in \mathbb{N}^N_{w+1 \leq |i| \leq w+N}} \prod_{m=1}^{N} t(i_m), \quad t(i_m) = \begin{cases} 1, & i_m = 1 \\ 2^{i_m-1} + 1, & i_m \geq 2. \end{cases}$$

The following bound has been shown in [12, Lemma 3.17]:

$$w \geq \frac{\log N_p}{\zeta + \log N} - 1$$

where $\zeta \approx 2.1$. Moreover in Theorem 4.2.8 we proved the following convergence rate for the sparse method

$$\hat{C}(r_{\min}, N) e^{-r_{\min} w}.$$ 

Combining these two results we can conclude that the error in terms of the number of collocation points is bounded by

$$\hat{C}(r_{\min}, N) e^{-r_{\min} \log N_p \log N}.$$ 

Therefore we expect an outcome with logarithmic decays for the error, but with a smaller deterioration rate. Namely the log of the error decays like $-\log N_p / \log N$.

We observe that for $N = 2$ the full tensor product method guarantees a slightly faster convergence than the sparse variant (Figure 5.1.5). On the other hand when $N$ becomes much bigger, the full tensor algorithm is affected by the curse of dimensionality.
We also examine the behaviour of the error related to the truncation of the Karhunen-Loève expansion of the diffusion coefficient $a$. In order to do this we fix both $p$ and $h$ and we study the convergence of the error obtained increasing the value of $N$. In practice we compare the approximation in the $N$-dimensional stochastic space to that one in the $N_{\text{max}}$-dimensional space, where $N_{\text{max}}$ is the maximum value of $N$ among those we consider in the implementation. The outcome is shown in Figure 5.1.6.

In the theoretical discussion we provided a bound for $\|u - u_N\|$ in case of a diffusion coefficient which depends linearly on the random variables. Nevertheless also in this exponential truncation case, we expect the error to decrease as $N$ increases. Rigorous rates of convergence should be investigated in future works.

Finally we are interested in comparing the results obtained by the stochastic collocation method to those coming from a Monte Carlo approach. The latter technique differs from the former for the choice of the evaluation points. Once we have fixed the number $K$ of realizations (which corresponds to the number of collocation points in the stochastic collocation method), we generate $K$ vectors $(y_k^{(MC)})_{k=1}^K$ of dimension $N$ uniformly distributed in $[-\sqrt{3}, \sqrt{3}]$ (which correspond to the collocation points). Given the evaluation points, we approximate by the finite element method the solutions to deterministic boundary value problems whose coefficients have been evaluated in each of the $K$ samples. This step is in common with the stochastic collocation approach. The expectation of the approximated solution $u_{N,K,h}$ is then recovered by computing the
Figure 5.1.6: The relative $L^2(D)$-norm error of the expectation of the truncated solution versus the dimension of the random domain $N$. Here $N_{\text{max}} = 9$ and $p_m = 2$ for all $m \in \{1, ..., N\}$.

The numerical results show that the convergence rate for the Monte Carlo algorithm is quite slow compared to that one we get adopting the stochastic collocation method, both in its full or sparse version. As a matter of fact we expect the Monte Carlo error to converge with an algebraic rate of $1/2$ with respect to the number of samples.

\[
\mathbb{E}_{\bar{\rho}}(u_{N,K,h})(x) = \frac{1}{K} \sum_{k=1}^{K} u_{N,h}(y_{k}^{(MC)}, x).
\]
Figure 5.1.7: The relative $L^2(D)$-norm error of the expectation of the final approximation computed by the Monte Carlo method and the full tensor product algorithm for $N = 2$.

Figure 5.1.8: The relative $L^2(D)$-norm error of the expectation of the final approximation computed by the Monte Carlo method and the sparse Smolyak algorithm for $N = 2$. 
5.2 Conclusions

The numerical examples presented in the previous section enlighten the effectiveness of the stochastic collocation method at least when the dimension $N$ of the random space (or equivalently the number of terms in the truncated Karhunen-Loève expansion of the input data) is quite small. This effectiveness combined with the easiness of implementation, which makes use of available deterministic codes, is the strength of the stochastic collocation technique. It outperforms the Monte Carlo method, both in its full tensor product construction and in the sparse tensor product one. Nevertheless the sparse version has to be preferred on the full method, especially for large values of $N$, as it reduces considerably the curse of dimensionality while keeping a high level of accuracy.

Unfortunately, the presented implementation is not capable of dealing with sparse grids of dimension $N > 2$. This limitation leaves open some important aspects like the validation of the effectiveness of the sparse tensor method in reducing the curse.

Further research has to be conducted also in the direction of anisotropic sparse tensor algorithms whose description has only been superficially treated in Section 3.2.1. Indeed if whereas the isotropic grid method works efficaciously with problems where the dependence on the random variables is equally weighted, the convergence rate gets worse when we try to approximate solutions of highly anisotropic problems which is the case when we consider Karhunen-Loève expansions of the coefficients (compare with Section 4.4). As a matter of fact in this case we have

$$\tau_m = \frac{1}{4\sqrt{\lambda_m} \| \phi_m \|_{L^\infty(D)}}$$

and therefore the analyticity region $\Sigma(\Gamma_m; \tau_m)$ increases as $m$ gets bigger.

Future works should also investigate the balance between the truncation error analysed in Section 4.2.1 and the collocation error extensively examined in the second half of Chapter 4. Indeed as we have shown the former error improves as $N$ increases, but the collocation error deteriorates in the same situation.
Summary

The purpose of this research project is to give an insight into the topic of the numerical approximation of elliptic partial differential equations with the property that the domain of the functions in the problem is the Cartesian product of a spatial and a probabilistic domain which we will denote as $D$ and $\Omega$, respectively. This results from modelling the uncertainties in a physical system. This approach sounds reasonable if we think at the unreliability and the numerous inaccuracies which pervade the real world. Mathematically we will deal with an equation taking the form

$$-\nabla \cdot (a(\omega, x)\nabla u(\omega, x)) = f(\omega, x), \quad (\omega, x) \in \Omega \times D. \quad (5.2.1)$$

Such a new kind of problem requires several tools needed for treating and eventually solving equation (5.2.1). In particular one of the main difficulties consists of the infinite dimensionality of the problem. Indeed whereas the spatial domain $D$ is usually a bounded subset of $\mathbb{R}^3$, in general the random domain $\Omega$ is infinite-dimensional. Thus the infinite-dimensional functions, which we model by the Karhunen-Loève expansion, have to be truncated in order to be manipulated on a finite memory machine. This truncation already introduces an error in our study.

A further step involves the actual resolution, or more realistically approximation, of the model equation (5.2.1) appended with some conditions at the boundary of the spatial domain $D$. At this stage we would like to exploit established techniques (for example the finite element scheme) which already allow to deal with deterministic elliptic problems, i.e. those where the domain is not probabilistic. In this direction the idea is to evaluate, with respect to the probabilistic variable, the coefficients of the problem at some suitably chosen points and subsequently to apply the finite element method which gives an approximation of the solution in the space $D$ for each of the points. The final approximation is recovered through the superposition of the spatial semi-approximations. This procedure is called the stochastic collocation method.

The approach is similar to the Monte Carlo algorithm but with a great difference in the selection of points. The evaluation grid of the stochastic collocation method is not random as in the case of the Monte Carlo technique but it consists of zeros of some suitable polynomials. This trick guarantees a much faster convergence in the error between the exact solution and the approximation. Some numerical examples at the end of the thesis aim to show the effectiveness of this theoretical result.
Bibliography


