Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs [*]

Albert Cohen, Ronald DeVore and Christoph Schwab

May 31, 2010

Abstract

Deterministic Galerkin approximations of a class of second order elliptic PDEs with random coefficients on a bounded domain $D \subset \mathbb{R}^d$ are introduced and their convergence rates are estimated. The approximations are based on expansions of the random diffusion coefficients in $L^2(D)$-orthogonal bases, and on viewing the coefficients of these expansions as random parameters $y = y(\omega) = (y_i(\omega))$. This yields an equivalent parametric deterministic PDE whose solution $u(x, y)$ is a function of both the space variable $x \in D$ and the in general countably many parameters y.

We establish new regularity theorems decribing the smoothness properties of the solution u as a map from $y \in U = (-1, 1)^\infty$ to $V = H_0^1(D)$. These results lead to analytic estimates on the V norms of the coefficients (which are functions of x) in a so-called "generalized polynomial chaos"(gpc) expansion of u.

Convergence estimates of approximations of u by best N-term truncated V-valued polynomials in the variable $y \in U$ are established. These estimates are of the form N^{-r}, where the rate of convergence r depends only on the decay of the random input expansion. It is shown that r exceeds the benchmark rate $1/2$ afforded by Monte-Carlo simulations with N "samples" (i.e. deterministic solves) under mild smoothness conditions on the random diffusion coefficients.

A class of fully discrete approximations is obtained by Galerkin approximation from a hierarchic family $\{V_l\}_{l=0}^\infty \subset V$ of finite element spaces in D of the coefficients in the N-term truncated gpc expansions of $u(x, y)$. In contrast to previous works, the level l of spatial resolution is adapted to the gpc coefficient. New regularity theorems decribing the smoothness properties of the solution u as a map from $y \in U = (-1, 1)^\infty$ to a smoothness space $W \subset V$ are established leading to analytic estimates on the W norms of the gpc coefficients and on their space discretization error. The space W coincides with $H^2(D) \cap H_0^1(D)$ in the case where D is a smooth or convex domain.

Our analysis shows that in realistic settings a convergence rate $N_{d.o.f}^{-s}$ in terms of the total number of degrees of freedom $N_{d.o.f}$ can be obtained. Here the rate s is determined by both the best N-term approximation rate r and the approximation order of the space discretization in D.

1

1   Introduction

Partial differential equations with stochastic input data are a commonly used paradigm in science and engineering. Stochasticity typically reflects the uncertainty in the various parameters entering the physical phenomenon described by the equation. A simple yet relevant model problem is the elliptic equation

$$-\nabla \cdot (a\nabla u) = f \quad \text{in} \quad D, \qquad u|_{\partial D} = 0, \tag{1.1}$$

in a bounded Lipschitz domain $D \subset \mathbb{R}^d$. We assume that $f = f(x)$ is a given deterministic function in $L^2(D)$ and that the diffusion coefficient a in (1.1) is a random field on a probability space (Ω, Σ, P) over $L^\infty(D)$ (see, e.g., [11]). In particular, given any $\psi \in L^2(D)$ and any Borel subset A of $\mathbb{R}$, the set $\{\omega \in \Omega : (a(\cdot,\omega), \psi) \in A\} \in \Sigma$ where $(\cdot,\cdot)_{L^2(D)}$ denotes the $L^2(D)$ innerproduct and $\omega \in \Omega$ represents a draw of this field with respect to the probability P.

In this model, stochasticity is therefore used to describe the uncertainty in the diffusion coefficient a. In order to ensure uniform ellipticity, we make the following assumption.

Assumption 1. *There exist constants* $0 < a_{\min} \leq a_{\max}$ *such that*

$$a_{\min} \leq a(x,\omega) \leq a_{\max}, \tag{1.2}$$

holds for all $(x,\omega) \in D \times \Omega$.

By the Lax-Milgram lemma, this assumption immediately implies for every $\omega \in \Omega$ the existence of a solution $u(\cdot, \omega)$ in the space $H_0^1(D)$, in the sense of the variational formulation:

$$\int_D a(x,\omega)\nabla u(x,\omega) \cdot \nabla v(x)dx = \int_D f(x)v(x)dx, \quad \text{for all} \ \ v \in H_0^1(D), \tag{1.3}$$

where the gradient ∇ is taken with respect to the x variable. This solution satisfies the estimate

$$\|u(\cdot,\omega)\|_V \leq B := \frac{\|f\|_{V^*}}{a_{\min}} \qquad \text{for all} \quad \omega \in \Omega. \tag{1.4}$$

Here, and in all the following, we denote by V the space $H_0^1(D)$, equipped with the energy norm $\|v\|_V := \|\nabla v\|_{L^2(D)}$ and by V^* its dual $H^{-1}(D)$. The estimate (1.4) also reads

$$\sup_{\omega \in \Omega} \|u(\cdot,\omega)\|_V \leq B. \tag{1.5}$$

The solution $u = u(x, \omega)$ is a random field associated to the probability space (Ω, Σ, P). Numerical methods have been developed in order to approximately compute quantities which describe the probabilistic behaviour of the field u. Such quantities are typically the *statistical moments* of u, such as

(i) the mean field $\bar{u}$ which is defined as (formal) "ensemble average"

$$\bar{u} := \mathbb{E}(u) = \int_\Omega u(\cdot, \omega) dP(\omega) \in V \ ,$$

(ii) the covariance function

$$C_u := \mathbb{E}([u - \bar{u}] \otimes [u - \bar{u}]) \ \in \ V \otimes V \ .$$

Since u is in general not a Gaussian process, $\bar{u}$ and C_u only give partial information on the probability distribution of u. We distinguish two general numerical approaches for computing such quantities: **Monte-Carlo (MC) methods** and **deterministic methods**.

Monte-Carlo methods: These are based on N independent draws $\{a_1, \cdots, a_N\}$ of the random coefficient a. For each instance a_i, they compute the solution u_i to the equation $-\nabla \cdot (a_i \nabla u_i) = f$ and use the resulting sample $\{u_1, \cdots, u_N\}$ to estimate the quantities of interest. For example, the mean field $\bar{u}$ is approximated by

$$\bar{u}_N := \frac{1}{N} \sum_{i=1}^{N} u_i. \tag{1.6}$$

The fact that the u_i are independent and their laws are identical to the law of u implies

$$\mathbb{E}(\|\bar{u} - \bar{u}_N\|_V^2) = \frac{1}{N} \mathbb{E}(\|u - \bar{u}\|_V^2)$$

and, since $\mathbb{E}(\|u - \bar{u}\|_V^2) \leq \mathbb{E}(\|u\|_V^2)$, we obtain with the Cauchy-Schwarz inequality,

$$\mathbb{E}(\|\bar{u} - \bar{u}_N\|_V) \leq (\mathbb{E}(\|u\|_V^2))^{1/2} N^{-\frac{1}{2}} \tag{1.7}$$

i.e. Monte-Carlo approximations with N samples converge with rate $1/2$ provided that the solution u as a V-valued random function has finite second moments. If u has lower summability, lower convergence rates for the MC approximation (1.6) will result: interpolating (1.7) with the straightforward bound $\mathbb{E}(\|\bar{u} - \bar{u}_N\|_V) \leq 2\mathbb{E}(\|u\|_V)$ implies the reduced rate N^{-r} with $r = 1 - 1/q \in [0, 1/2]$ provided $\mathbb{E}(\|u\|_V^q) < \infty$ for some $q \in [1, 2]$ (see e.g. [14]). Better summability of u in the ω variable, such as e.g. (1.5), however, does *not* generally allow to improve the convergence rate of the MC approximation (1.6) beyond $r = 1/2$.

In practice, the u_i in (1.6) are computed approximately by space discretization, for example by the finite element method. The computable approximation to $\bar{u}$ is thus given by

$$\bar{u}_{N,h} := \frac{1}{N} \sum_{i=1}^{N} u_{i,h},$$

3

where $u_{i,h}$ is the Galerkin approximation to u_i in the finite element space V_h which need to be chosen such that the corresponding discretization error does not affect the MC rate (1.7). The total complexity of the solution process is thus at least of order $\mathcal{O}(NM)$ where $M = \dim(V_h)$.

Leaving aside the space discretization aspects, we note the following advantages and drawbacks of the MC approximation (1.6):

- On the positive side, the computations of the u_i are independent from each other and can be performed in a parallel fashion. Observe also that MC is a statistical inference approach: it does not require the full knowledge of the joint probability law of the field a, but only a sample of independent instances. If, however, these instances are computer generated (as, e.g., in numerical simulations), the "simulator" necessarily contains the law of a in some (parametric) form.

- On the negative side, the convergence estimate (1.7) is only in an expectation sense (although $\overline{u}$ is a deterministic function).

  The convergence rate (1.7) of $1/2$ for the Monte-Carlo approximation (1.6) can not be improved, in general, despite the fact that the solution u depends smoothly on a.

Deterministic methods: These have been studied for several decades (see [8] and the references therein). In contrast to MC, these methods take advantage of the smooth dependence of u on a. We distinguish two general classes of deterministic methods.

The *perturbation approach* is based on the Neumann expansion of the stochastic solution around its mean field, and successive computations of the terms in this expansion (see [10] and the references therein). Such methods are computationally efficient for the first terms, i.e. the low order moments of the solution, yet grow in complexity for higher order terms.

The *spectral approach* is based on the so-called Wiener/generalized polynomial chaos expansion introduced in [22] (see also [9] and [15]). The first step consists in representing a by a sequence of scalar random variables $(y_j)_{j \geq 1}$, usually obtained through a decomposition of the oscillation $a - \overline{a}$ into an orthogonal basis $(\psi_j)_{j \geq 1}$ of $L^2(D)$:

$$a(x, \omega) = \overline{a}(x) + \sum_{j \geq 1} y_j(\omega) \psi_j(x). \tag{1.8}$$

The solution is now viewed as a function $u(x, y)$ where $x \in D$ is the space variable and $y = (y_j)_{j \geq 1}$ is a vector of "stochastic variables", and the objective is to compute a numerical approximation to $u(x, y)$. This approach provides an approximation of the probability law of the solution and therefore gives access to virtually all possible information on its probabilistic behaviour. However, one is facing a problem of high - possibly infinite - dimension, due to the number of coordinates in the y variable refers to the approximation of the solution in this variable using tensor product polynomials.

The numerical analysis of the spectral approach began only recently. When the vector y has finite dimension K, the error between $u(x, y)$ and its approximation were shown in [1] to decrease exponentially with respect to the polynomial degree of the approximation. However, the derived estimates depend heavily on K as K grows. Since y is usually of infinite dimension,

one needs to incorporate the effect of its truncation to a finite set of K variables in the error analysis, and K has to grow to $+\infty$ in the convergence analysis. There is therefore a crucial need for convergence estimates which are *independent* of K. Such estimates were established for the first time in [21], where exponential convergence rates independent of K were proved under the assumption that the terms in the expansion (1.8) decay exponentially to 0 in the L^∞ norm. Here, the authors specialized on the Karhúnen-Loève (KL) expansion for (1.8), and the desired decay property was obtained under the (strong) assumption that the covariance function

$$C_a(x,y) := \mathbb{E}([a(x) - \bar{a}(x)][a(y) - \bar{a}(y)]) \tag{1.9}$$

has analytic smoothness in x and y.

In the present paper, we explore the more realistic setting where the terms in (1.8) only have algebraic decay. Our main objective is to design an approximation scheme which converges with rate $r > 1/2$ under such realistic assumptions on the random input. Our analysis is not restricted to the KL expansion and will therefore be carried out for general expansions of a in an orthogonal basis (ψ_j). We shall analyze the dependence of the solution $u(x,y)$ on the parameters y and thereby show that u has an expansion into a polynomial basis with coefficients from V. By deriving a priori bounds on the decay of the coefficients of u in such an expansion, we shall derive algebraic rates of convergence for the spectral approach under rather mild assumptions on the smoothness of a. Our analysis is independent of the number K of retained variables and depends only on the rate of decay of the terms in (1.8). A key feature in our analysis lies in the choice of a particular sparse tensor product polynomial space, which can be interpreted as a form of non-linear best N-term approximation. Let us mention that other strategies for selecting the sparse polynomial spaces in the variable $y \in U$ have been proposed and investigated recently in [12, 13] based on the concept of sparse grid introduced in [19]. A specific feature of our approach, compared to these strategies, lies in the optimal choice of the polynomial space which allows us to relate the convergence rate r to the rate of decay of the random input expansion. Another contrast with previous works is that we adapt the level of spatial discretization to each gpc coefficient, which is essential in order to obtain an optimal overall convergence rate in terms of the total number of degrees of freedom.

Our paper is organized as follows. We discuss in §2 the general properties of the stochastic expansion (1.8) and introduce the corresponding parametric PDE induced by (1.1). This parametric problem is defined for parameters $y \in U$ where U is the set of all sequences (y_j) with $|y_j| \leq 1$. In §3, we introduce the measures and spaces defined on U which are the setting for our approach. This is followed in §4 by deriving bounds on the partial derivatives of $u(x,y)$ with respect to the variables y_j. A general Galerkin scheme for the approximation of $u(x,y)$ in the y variable is proposed in §5, based on a sparse set of tensorized Legendre polynomials. In order to study the convergence of this scheme, we investigate in §6 bounds on the exact Legendre coefficients in the expansion of u. These estimates are used in order to derive the convergence rate of the Galerkin scheme, through several key results on the summability of multi-indexed sequences which are established in §7. Finally we discuss in §8 the full discretization in the x and y variables and make a final comparison with MC methods. Conclusions and perspectives are raised in the final section.

We emphasize that the results of this paper show that solutions to stochastic and parametric equations of the above type possess enough regularity to be well approximated by Galerkin subspaces of suitable dimension. The problem of how to numerically find these subspaces is not treated although we make some remarks on this interesting problem in the concluding section of this paper.

2   Basis expansions of the coefficient a

The present paper is based on the spectral approach which we recall begins by decomposing the random field a into an expansion of the type (1.8). We assume throughout this paper that $(\psi_j)_{j \geq 1}$ is a complete orthogonal sequence in $L^2(D)$ (we could work more generally with any Riesz basis of $L^2(D)$). Based on our assumptions on the random field a, the random variables

$$y_j := \|\psi_j\|_{L^2}^{-2} \int_D (a - \bar{a})\psi_j, \quad j = 1, 2, \dots$$

are P-measurable functions.

We next introduce assumptions concerning the convergence of the expansion (1.8) in the $L^\infty(D)$ norm. These assumptions are formulated in terms of the summability properties of the sequence $(\|y_j\psi_j\|_{L^\infty(D)})_{j\geq1}$. Up to a renormalization of the basis functions ψ_j, we may assume without loss of generality that for all $j \geq 1$ the random variables y_j are such that $\|y_j\|_{L^\infty(\Omega)} = 1$. Up to a change of the definition of a on a set of measure zero in Ω this is equivalent to

$$\sup_{\omega \in \Omega} |y_j(\omega)| = 1. \tag{2.1}$$

The vector y is thus supported in the infinite dimensional cube

$$U := [-1, 1]^{\mathbb{N}},$$

i.e. the unit ball of $\ell^\infty(\mathbb{N})$. With such a normalization, our assumptions are formulated on the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq1}$. Our first assumption is a strengthening of Assumption 1.

Assumption 2. *The functions $\bar{a}$ and ψ_j satisfy*

$$\sum_{j\geq1} \|\psi_j\|_{L^\infty(D)} \leq \frac{\kappa}{1 + \kappa}\bar{a}_{\min}, \tag{2.2}$$

with $\bar{a}_{\min} := \min_{x \in D} \bar{a}(x) > 0$ and $\kappa > 0$.

In the convergence results of §7, a prescribed value of the constant κ will be needed. We can view Assumption 2 as a strong ellipticity assumption on a which requires that the relative perturbation of $\bar{a}$ by the series $\sum_{j\geq1} y_j\psi_j$ is not too large. Clearly, it implies Assumption 1 with

$$a_{\min} := \bar{a}_{\min} - \frac{\kappa}{1 + \kappa}\bar{a}_{\min} = \frac{1}{1 + \kappa}\bar{a}_{\min} > 0. \tag{2.3}$$

Since $\frac{\kappa}{1+\kappa}\overline{a}_{\min} = \kappa a_{\min}$, Assumption 2 also implies

$$\sum_{j \geq 1} \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}} < \kappa. \tag{2.4}$$

In particular, Assumption 2 and the value of κ is independent of the scaling of a.

In order to obtain convergence rates $r > \frac{1}{2}$ for our approximation scheme, additional summability properties are needed as expressed by the following assumption.

Assumption 3. *The sequence* $(\|\psi_j\|_{L^\infty(D)})$ *belongs to* $\ell^p(\mathbb{N})$ *for some* $p < 1$:

$$\sum_{j \geq 1} \|\psi_j\|_{L^\infty(D)}^p < +\infty$$

We next discuss possible choices for the basis $(\psi_j)_{j \geq 1}$. Since the main objective is to describe accurately the diffusion coefficient a with as few parameters y_j as possible and to fullfill the above summability assumptions, this choice should be tied to the properties of this random field. On the other hand this basis will enter the computation of the solution and should therefore be easy to manipulate numerically.

An important example is the Karhúnen-Loève basis of the $L^2(D)$-orthogonal eigenfunctions of the covariance integral operator

$$g \mapsto Tg(x) := \int_D C_a(x,y)g(y)dy,$$

where C_a is the covariance function (1.9). We index these eigenfunctions in decreasing order of the corresponding eigenvalues. These functions are well defined for any domain D. In the particular case where D is a fundamental period, $D = [0,1]^d$ say, and a is a stationary and D-periodic random field, i.e. its covariance function has the form $C_a(x,y) = A(x-y)$ where A is D-periodic, then T is a convolution operator and the KL basis is the Fourier basis. In general, the KL expansion has properties which emulate those of Fourier series. In particular, the decay of the terms and the rate of convergence of the KL expansion are dictated by the average regularity of the field a measured by the smoothness properties of C_a. We refer to [21] for such results when C_a is analytic and to [20] for similar results with less regular kernels.

In the case of a one-dimensional Fourier expansion,

$$a(x,\omega) = \overline{a}(x) + \sum_{k \in \mathbb{Z}} \hat{a}(k,\omega)e^{i2\pi kx},$$

it is known that if the function $a(\cdot,\omega) - \overline{a}$ is in $\text{Lip}(s, L^1)$ for some $s > 1$, then its Fourier coefficients satisfy the decay estimate

$$|a(k,\omega)| \leq C|k|^{-s}, \qquad |k| \geq 1,$$

with C depending on the Lip(s, L^1)-norm of $a(\cdot, \omega) - \bar{a}$. Assuming that this norm is bounded independently of ω and reindexing the expansion on $j \geq 1$ with the normalization (2.1), we thus obtain

$$\|\psi_j\|_{L^\infty(D)} \leq C j^{-s}, \qquad j = 1, 2, \dots \tag{2.5}$$

Therefore ℓ^p summability of the sequence $(\|\psi_j\|_{L^\infty(D)})_{j \geq 1}$ is ensured when $s > \frac{1}{p}$. In the multivariate case, the rate of decay (2.5) is modified to $j^{-s/d}$. In summary, Assumption 2 and 3 can be derived from the smoothness properties on the field a.

In the nonperiodic case, the KL eigenfunctions are in general not analytically available. However, approximations of them can be computed efficiently numerically; see, e.g. [18] for algorithms based on fast multipole approximations of covariance operators T when $C_a(x, y)$ is analytic for $x \neq y$. There exist, however, many other basis expansions for which similar decay properties hold when a has some smoothness, in particular wavelet expansions which can be constructed on fairly arbitrary domains, see [4] for a general treatment. One can carry out for wavelet bases, the same analysis as described above for the Fourier basis. For example, in the univariate case, the decay rate (2.5) is now satisfied if $a(\cdot, \omega) - \bar{a}$ belong to the Hölder space C^s with their C^s-norm bounded independently of ω.

In summary, the basis ψ_j should be taken with an eye towards two issues. The first is that it should be easy to manipulate numerically. The second is that the infinite dimensional vector y has components y_j which decrease rapidly as j grows, with the rate of decay being determined by the smoothness of the field a.

For each $y \in U$, we define

$$a = a(x, y) := \bar{a} + \sum_{j \geq 1} y_j \psi_j(x), \quad x \in D, y \in U. \tag{2.6}$$

Because of Assumption 2 , the series (2.6) converges absolutely and uniformly on $D \times U$. Notice that $a(x, y)$ is defined for all $y \in U$ and not just for the $y(\omega)$ which are the image of some $\omega \in \Omega$. In particular, we have

$$a(x, y) \geq a_{\min}, \tag{2.7}$$

for all $y \in U$ with $a_{\min}$ defined by (2.3). In the sequel, we will use a to denote both $a(x, \omega)$ and $a(x, y)$ but which of these is being employed will be clear from the context. Similarly y will denote both the stochastic basis coefficients $y(\omega)$ as well as a general point in the parameter space U.

3   Probability spaces on U

Since U is an infinite product of the intervals $[-1, 1]$ some care must be taken in defining probability measures on U. We shall have need for two measures. The first of these is the infinite tensor product measure $d\mu$ of the univariate uniform probability measures on $[-1, 1]$:

$$d\mu(y) = \otimes_{j \geq 1} dy_j / 2. \tag{3.1}$$

Recall that the sigma algebra for $d\mu$ is generated by the finite rectangles $\prod_{j=1}^{\infty} S_j$, where only a finite number of the S_j are different from $[-1,1]$ and those that are different are intervals contained in $[-1,1]$. Then $(U, \Theta, d\mu)$ is a probability space.

We shall also need a measure ρ defined on U which is induced by the mapping $\omega \to y(\omega)$. This measure is defined on the same sigma algebra Θ as for the uniform measure discussed above. Consider any finite rectangle $\otimes_{j=1}^{\infty} S_j$, where $S_j = [-1,1]$ for all $j \geq n$ for some n. We define

$$\rho(S) := \prod_{j=1}^{n} P\{\omega : y_j(\omega) \in S_j\}. \tag{3.2}$$

Then, ρ extends to a measure defined on all sets in the sigma algebra Θ. This gives the measure space (U, Θ, ρ). Given these measure spaces, we introduce for $1 \leq p \leq \infty$ the Banach spaces $L^p(U, d\mu)$ and $L^p(U, d\rho)$. For the (separable) Hilbert space V, we denote by $L^p(U, V, d\mu)$ and $L^p(U, V, d\rho)$ the corresponding Bochner spaces of p-summable mappings from U to V, equipped with their corresponding norms. For example,

$$\|v\|_{L^2(U,V,d\rho)}^2 := \int_U \|v(\cdot, y)\|_V^2 d\rho(y) = \int_U \left(\int_D |\nabla v(x,y)|^2 dx \right) d\rho(y). \tag{3.3}$$

Here and in the following, ∇ is understood to be applied in the x variable.

We shall also need certain orthogonal bases, built from Legendre polynomials, for some of these spaces. Let $(L_n)_{n \geq 0}$ be the univariate Legendre polynomials normalized according to

$$\int_{-1}^{1} |L_n(t)|^2 \frac{dt}{2} = 1, \tag{3.4}$$

or equivalently

$$\max_{t \in [-1,1]} |L_n(t)| = \sqrt{2n+1}. \tag{3.5}$$

Recall that L_n is an algebraic polynomial of degree n and the family $(L_n)_{n \geq 0}$ is a complete orthogonal system for $L^2[-1,1]$.

We introduce the countable set $\mathcal{F}$ of all sequences $\nu = (\nu_j)_{j \geq 1}$ of nonnegative integers such that only finitely many ν_j are non-zero. For all $\nu \in \mathcal{F}$, we use the notation

$$|\nu| := \sum_{j \geq 1} \nu_j = \|\nu\|_{\ell^1},$$

and

$$\nu! = \prod_{j \geq 1} \nu_j!, \quad \nu \in \mathcal{F}.$$

We define the tensorized Legendre polynomials by

$$L_\nu(y) = \prod_{j \geq 1} L_{\nu_j}(y_j). \tag{3.6}$$

By construction, the family $(L_\nu)_{\nu \in \mathcal{F}}$ forms an orthonormal system in $L^2(U, d\mu)$. Since $L_0(t) = 1$, and any $\nu \in \mathcal{F}$ has only a finite number of nonzero entries, the function $L_\nu(y)$ only depends on finitely many y_j (namely those j such that $\nu_j \neq 0$).

9

The family $(L_\nu)_{\nu \in \mathcal{F}}$ is easily seen to be complete in $L^2(U, d\mu)$. Indeed, any function in $L^2(U, d\mu)$ can be approximated to any given tolerance by a finite linear combination of characteristic functions of finite rectangles and each characteristic function of a finite rectangle can be approximated by polynomials to any prescribed accuracy. Therefore $(L_\nu)_{\nu \in \mathcal{F}}$ is an orthonormal basis of $L^2(U, d\mu)$. In turn, each $v \in L^2(U, V, d\mu)$ has a representation

$$v = \sum_{\nu \in \mathcal{F}} v_\nu L_\nu, \quad \text{where} \quad v_\nu = \int_U g(\cdot, y) L_\nu(y) d\mu(y) \in V \tag{3.7}$$

and $\|v\|_{L^2(U,V,d\mu)} = \|(\|v_\nu\|_V)\|_{\ell^2(\mathcal{F})}$.

4   Parametric expansion of u

Suppose that we have an orthogonal system $(\psi_j)_{j \geq 0}$ such that Assumption 2 holds and a is defined by (2.6). We denote by $u(x, y)$ the solution to

$$-\nabla \cdot (a\nabla u) = f \quad \text{in} \quad D, \qquad u|_{\partial D} = 0, \tag{4.1}$$

where $D \subset \mathbb{R}^d$ is the Lipschitz domain introduced earlier. Since Assumption 2 implies the lower bound (2.7), the equations (4.1) are uniformly elliptic in $y \in U$, and we have

$$\|u\|_{L^\infty(U,V)} = \|u\|_{L^\infty(U,V;d\mu)} := \sup_{y \in U} \|u(\cdot, y)\|_V \leq B \tag{4.2}$$

with B as in (1.4). Throughout in what follows, the expression $L^\infty(U, V)$ shall be understood in the sense (4.2), also for different choices of the space V.

In this section, we fix f and examine the smoothness of u as a function of the parameter vector $y \in U$. We shall establish generic a-priori bounds for $\|\partial_y^\nu u\|_{L^\infty(U,V)}$. These bounds could possibly be improved when working with a specific orthogonal system $(\psi_j)_{j \geq 1}$ such as a wavelet system. However, at this stage we do not want to complicate the presentation by pursuing such avenues.

As a first step, we prove the existence in V of the partial derivatives $\partial_y^\nu u$ at any $y \in U$. For this purpose, we need the following stability result.

Lemma 4.1 *If u and $\tilde{u}$ are solutions of (1.3) with the same right hand side f and with coefficients a and $\tilde{a}$, respectively, and if these coefficients both satisfy the assumption (1.2) with the same lower bound $a_{\min}$, then*

$$\|u - \tilde{u}\|_V \leq \frac{\|f\|_{V^*}}{a_{\min}^2} \|a - \tilde{a}\|_{L^\infty(D)}. \tag{4.3}$$

Proof: Substracting the variational formulations (1.3) for u and $\tilde{u}$, we find that for all $v \in V$,

$$0 = \int_D a\nabla u \cdot \nabla v - \int_D \tilde{a}\nabla \tilde{u} \cdot \nabla v = \int_D a(\nabla u - \nabla \tilde{u}) \cdot \nabla v + \int_D (a - \tilde{a})\nabla \tilde{u} \cdot \nabla v. \tag{4.4}$$

10

Therefore $w = u - \tilde{u}$ is the solution of $\int_D a\nabla w \cdot \nabla v = L(v)$ where $L(v) := \int_D (a - \tilde{a})\nabla\tilde{u} \cdot \nabla v$. Hence

$$\|w\|_V \leq \frac{\|L\|_{V^*}}{a_{\min}},$$

and we obtain (4.3) since

$$\|L\|_{V^*} = \max_{\|v\|_V = 1} |L(v)| \leq \|a - \tilde{a}\|_{L^\infty(D)}\|\tilde{u}\|_V \leq \|a - \tilde{a}\|_{L^\infty(D)}\frac{\|f\|_{V^*}}{a_{\min}}.$$

$\diamond$

Theorem 4.2 *At any $y \in U$, the function $y \mapsto u(y)$ admits a partial derivative $\partial_y^\nu u(y) \in V$ for any $\nu \in \mathcal{F}$.*

Proof: We start by proving the existence of the first order derivative $\partial_{y_j}^\nu u(y)$ for any $j \geq 1$ and $y \in U$. We denote by e_j the Kronecker sequence with 1 at index j and 0 at other indices. For $h \in \mathbb{R} \setminus \{0\}$ we consider the difference quotient

$$w_h(y) = \frac{u(y + he_j) - u(y)}{h}. \tag{4.5}$$

We notice that this quotient is well defined for h small enough: if $|h|\|\psi_j\|_{L^\infty(D)} \leq \frac{a_{\min}}{2}$, we clearly have for all $y \in U$,

$$\frac{a_{\min}}{2} \leq a(x, y + he_j) \leq a_{\max} + \frac{a_{\min}}{2}, \quad x \in D,$$

and therefore $u(y + he_j)$ is well defined as an element of V. For such a small enough h, we have for all $v \in V$,

$$
\begin{aligned}
0 &= \int_D a(x, y + he_j)\nabla u(x, y + he_j) \cdot \nabla v(x)dx - \int_D a(x, y)\nabla u(x, y) \cdot \nabla v(x)dx \\
&= h\int_D a(x, y)\nabla w_h(x, y) \cdot \nabla v(x)dx + \int_D (a(x, y + he_j) - a(x, y))\nabla u(x, y + he_j) \cdot \nabla v(x)dx \\
&= h\int_D a(x, y)\nabla w_h(x, y) \cdot \nabla v(x)dx + h\int_D \psi_j(x)\nabla u(x, y + he_j) \cdot \nabla v(x)dx
\end{aligned}
$$

and therefore $w_h(y)$ is the unique solution to

$$\int_D a(x, y)\nabla w_h(x, y) \cdot \nabla v(x)dx = L_h(v), \quad \text{for all } v \in V,$$

where $L_h : v \to L_h(v) := -\int_D \psi_j(x)\nabla u(x, y + he_j) \cdot \nabla v(x)dx$ is a continuous, linear functional on V. The linear functional $L_h(\cdot)$ varies continuously in V^* with h as h tends to 0: indeed, we have for all $v \in V$,

$$|L_h(v) - L_0(v)| = |\int_D \psi_j(\nabla u(y + he_j) - \nabla u(y)) \cdot \nabla v| \leq \|\psi_j\|_{L^\infty(D)}\|u(y + he_j) - u(y)\|_V\|v\|_V,$$

and since the stability estimate (4.3) implies

$$\|u(y + he_j) - u(y)\|_V = \|\nabla u(y + he_j) - \nabla u(y)\|_{L^2(D)} \leq |h|\|\psi_j\|_{L^\infty(D)}\frac{4\|f\|_{V^*}}{a_{\min}^2},$$

11

it follows that L_h converges towards L_0 in V^* as $h \to 0$. Therefore w_h converges in V towards w_0, which is the solution to

$$\int_D a(x,y)\nabla w_0(x) \cdot \nabla v(x)dx = L_0(v), \quad \text{for all } v \in V.$$

Hence $\partial_{y_j} u(y) = w_0$ exists in V and is the unique solution of the variational problem

$$\int_D a(x,y)\nabla \partial_{y_j} u(x,y)\nabla v(x)dx = -\int_D \psi_j(x)\nabla u(x,y)\nabla v(x)dx, \quad \text{for all } v \in V. \tag{4.6}$$

We notice that this problem is obtained by formally differentiating the variational formulation: given $f \in V^*$ and $y \in U$, find $u(\cdot, y) \in V$ such that

$$\int_D a(x,y)\nabla u(x,y)\nabla v(x)dx = \int_D f(x)v(x)dx , \quad \text{for all } v \in V \tag{4.7}$$

with respect to the variable y_j. By the same reasoning, we can inductively derive the existence in V of higher order derivatives $\partial_y^\nu u(y)$. These derivatives are solutions of variational problems obtained by further differentiating (4.7). $\diamond$

We now estimate the norms $\|\partial_y^\nu u\|_{L^\infty(U,V)}$, For this purpose, we introduce the following notation. If $\alpha = (\alpha_j)_{j\geq 1}$ is a sequence of positive numbers, we define for all $\nu \in \mathcal{F}$

$$\alpha^\nu := \prod_{j\geq 1} \alpha_j^{\nu_j}.$$

We also use the following sequence b throughout the remainder of this paper:

$$b = (b_j)_{j=1}^\infty, \qquad b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}. \tag{4.8}$$

Theorem 4.3 *With the constant B as in (1.4), we have*

$$\|\partial_y^\nu u\|_{L^\infty(U,V)} \leq B|\nu|!b^\nu \quad \forall \nu \in \mathcal{F}. \tag{4.9}$$

Proof: As a first step, we study the variational problems satisfied by the partial derivatives $\partial_y^\nu(y)$. We claim that these problems have the general recursive form

$$\int_D a(x,y)\nabla \partial_y^\nu u(x,y)\nabla v(x)dx = - \sum_{\{j:\, \nu_j \neq 0\}} \nu_j \int_D \psi_j(x)\nabla \partial_y^{\nu-e_j} u(x,y)\nabla v(x)dx, \tag{4.10}$$

where e_j is again the Kronecker sequence with value 1 at position j and 0 elsewhere.

We prove (4.10) by induction on $|\nu|$. When $|\nu| = 1$ this is (4.6). For $|\nu| > 1$, let k be any index such that $\nu_k \neq 0$, we define $\tilde{\nu} = \nu - e_k$ which satisfies $|\tilde{\nu}| = |\nu| - 1$. By the induction hypothesis, we have for all $v \in V$

$$\int_D a(x,y)\nabla \partial_y^{\tilde{\nu}} u(x,y)\nabla v(x)dx + \sum_{\{j:\, \tilde{\nu}_j \neq 0\}} \tilde{\nu}_j \int_D \psi_j(x)\nabla \partial_y^{\tilde{\nu}-e_j} u(x,y)\nabla v(x)dx = 0,$$

12

where $\tilde{\nu}_j = \nu_j$ if $j \neq k$ and $\tilde{\nu}_k = \nu_k - 1$. Differentiating with respect to y_k, we obtain

$$0 = \int_D a(x,y)\nabla\partial_y^\nu u(x,y)\nabla v(x)dx + \int_D \psi_k(x)\nabla\partial_y^{\nu-e_k}u(x,y)\nabla v(x)dx$$
$$+ \sum_{\{j\neq k:\, \nu_j\neq 0\}} \nu_j \int_D \psi_j(x)\nabla\partial_y^{\nu-e_j}u(x,y)\nabla v(x)dx + (\nu_k-1)\int_D \psi_k(x)\nabla\partial_y^{\nu-e_k}u(x,y)\nabla v(x)dx,$$

which is equivalent to (4.10). Selecting in (4.10) the function $v(x) = \partial_y^\nu u(x,y) \in V$, and using both ellipticity and continuity of the bilinear form, we obtain

$$a_{\min}\|\partial_y^\nu u(\cdot,y)\|_V^2 \le \int_D a(x,y)|\nabla\partial_y^\nu u(x,y)|^2 dx$$
$$= -\sum_{\{j:\nu_j\neq 0\}} \nu_j \int_D \psi_j(x)\nabla\partial_y^{\nu-e_j}u(x,y)\nabla\partial_y^\nu u(x,y)dx$$
$$\le \sum_{\{j:\nu_j\neq 0\}} \nu_j \|\psi_j\|_{L^\infty(D)}\|\partial_y^\nu u(\cdot,y)\|_V\|\partial_y^{\nu-e_j}u(\cdot,y)\|_V,$$

and therefore

$$\|\partial_y^\nu u(\cdot,y)\|_V \le \sum_{\{j:\nu_j\neq 0\}} \nu_j b_j \|\partial_y^{\nu-e_j}u(\cdot,y)\|_V. \tag{4.11}$$

Using (4.11), we now prove (4.9) by induction on $|\nu|$. For $|\nu| = 0$ this bound is simply $\|u(\cdot,y)\|_V \le B$ with B as in (1.4) which is known from (4.2). For $|\nu| > 0$, we combine (4.11) with the induction hypothesis. This yields

$$\|\partial_y^\nu u(\cdot,y)\|_V \le B \sum_{\{j:\, \nu_j\neq 0\}} \nu_j b_j (|\nu|-1)! b^{\nu-e_j} = B\Big(\sum_{\{j:\, \nu_j\neq 0\}} \nu_j\Big)(|\nu|-1)! b^\nu = B|\nu|! b^\nu,$$

which concludes the proof. ◇

5　Galerkin approximation

In this section, we shall introduce a numerical approach for the computation of $u(x,y)$. We assume that we have full knowledge of $d\rho$ (which may or may not be the case in a given application). Obviously u belongs to $L^2(U,V,d\rho)$ (since $\|u(\cdot,y)\|_V$ is uniformly bounded with respect to $y \in U$) and it can be defined as the unique solution of the variational problem:

$$\text{Find } u \in L^2(U,V,d\rho) \quad \text{such that} \quad B(u,v) = F(v) \quad \forall v \in L^2(U,V,d\rho), \tag{5.1}$$

where

$$B(u,v) := \int_U \Big(\int_D a(x,y)\nabla u(x,y)\cdot\nabla v(x,y)dx\Big)d\rho(y) \text{ and } F(v) := \int_U \Big(\int_D f(x)v(x,y)dx\Big)d\rho(y). \tag{5.2}$$

For any subset $\Lambda \subset \mathcal{F}$ of finite cardinality, we define the approximation space

$$X_\Lambda := \{v_\Lambda(x,y) = \sum_{\nu\in\Lambda} v_\nu(x)L_\nu(y)\,;\, v_\nu \in V\},$$

13

where $\{L_\nu\}_{\nu \in \mathcal{F}}$ is the basis of Legendre polynomials. Note that $X_\Lambda \subset L^\infty(U, V) \subset L^2(U, V, d\rho)$. We define the *Galerkin approximation* $u_\Lambda = \sum_{\nu \in \Lambda} u_\nu L_\nu \in X_\Lambda$ to u as the unique solution to the problem: find

$$u_\Lambda \in X_\Lambda \quad \text{such that} \quad B(u_\Lambda, v_\Lambda) = F(v_\Lambda) \quad \forall v_\Lambda \in X_\Lambda. \tag{5.3}$$

Just as for the MC method, the evaluation of u_Λ requires the computation of N deterministic functions u_ν where $N := \#(\Lambda)$, and the computation of these functions requires in addition some spatial discretization. We postpone discussion of the spatial discretization until §6 and first focus our analysis on the discretization in the y variable. Namely, we search for appropriate choices of Λ with the goal of obtaining error estimates of the form

$$\|u - u_\Lambda\|_{L^2(U,V,d\rho)} \leq C N^{-r}, \tag{5.4}$$

for the largest possible $r > 0$.

Remark 5.1 *We can derive from u_Λ an approximation to the mean field $\bar{u} = \mathbb{E}(u)$ which are given by*

$$\bar{u}_\Lambda = \mathbb{E}(u_\Lambda) = \sum_{\nu \in \Lambda} e_\nu u_\nu, \tag{5.5}$$

with the ν-th moments $e_\nu := \mathbb{E}(L_\nu(y)) = \int_U L_\nu(y)d\rho(y)$ (although the mean here is taken with respect to y and the measure $d\rho$ it is easily seen that it results in the same means $\bar{u}$ as averaging with respect to ω and P) . By the triangle and Cauchy-Schwarz inequalities,

$$\|\bar{u} - \bar{u}_\Lambda\|_V \leq \int_U \|u(\cdot, y) - u_\Lambda(\cdot, y)\|_V d\rho(y) \leq \|u - u_\Lambda\|_{L^2(U,V,d\rho)}. \tag{5.6}$$

Therefore the rate of the spectral Galerkin approximation (5.3) will outperform the rate (1.7) of the MC estimate (1.6) for the mean field $\mathbb{E}(u)$ in terms of the number N of coefficients in V to be determined, if $r > \frac{1}{2}$ in (5.4).

Remark 5.2 *Our approach implicitly assumes that we have the full knowledge of ρ or equivalently of P, in contrast to the MC method which only needs a sample of independent realizations. In the case where we only have such a sample $(y^1, \cdots, y^M) \in U^M$ at our disposal, we can adapt our approach by solving*

$$B_M(u_\Lambda, v_\Lambda) = F_M(v_\Lambda), \tag{5.7}$$

in place of (5.3), where B_M and F_M are defined by replacing the integrals of the type $\int f(y)d\rho(y)$ in (5.2) by their empirical counterpart $\frac{1}{M} \sum_{i=1}^M f(y^i)$. We shall not embark in the error analysis of this variant and proceed with the assumption that ρ is known to us.

Remark 5.3 *An alternative to Galerkin discretization would have been to start from an orthonormal basis of $L^2(U, d\rho)$ instead of $L^2(U, d\mu)$. However such a basis is not always simple to construct when ρ is not separable and therefore we maintain the choice of the Legendre polynomials even when ρ differs from μ. As we shall see later, sharper error estimates can be obtained in the particular case where $\rho = \mu$, i.e. when the random variables $y_j(\omega)$ are independent and uniformly distributed on $[-1, 1]$.*

Remark 5.4 *An alternate approach to Galerkin discretization is collocation: find $u_\Lambda \in X_\Lambda$ such that*

$$\int_D a(x,y)\nabla u_\Lambda(x,y)\nabla v(x)dx = \int_D f(x)v(x)dx, \tag{5.8}$$

for all $v \in V$ and for all $y \in S_\Lambda$ where $S_\Lambda \subset U$ is a set such that $\#(S_\Lambda) = N$. This approach is however more difficult to analyze, since its well-posedness is strongly tied to an optimal choice of S_λ.

We begin our error analysis by observing that according to Cea's lemma (see e.g. [3]), we have the estimate

$$\|u - u_\Lambda\|_{L^2(U,V,d\rho)} \leq C_1 \inf_{v_\Lambda \in X_\Lambda} \|u - v_\Lambda\|_{L^2(U,V,d\rho)}, \tag{5.9}$$

where $C_1 := \sqrt{\frac{a_{\max}}{a_{\min}}}$. In order to proceed further, we introduce the exact expansion of u in the basis $(L_\nu)_{\nu \in \mathcal{F}}$ (see (3.7)):

$$u(x,y) = \sum_{\nu \in \mathcal{F}} c_\nu(x)L_\nu(y) \quad \text{where} \quad c_\nu := \int_U u(\cdot,y)L_\nu(y)d\mu(y) \in V.$$

We infer from (5.9) that

$$\|u - u_\Lambda\|_{L^2(U,V,d\rho)} \leq C_1 \|u - \sum_{\nu \in \Lambda} c_\nu L_\nu\|_{L^2(U,V,d\rho)}. \tag{5.10}$$

The right hand side of (5.10) can be bounded in different ways depending on the properties of ρ with respect to μ.

- **Case 1:** if $\rho = \mu$, we can invoke Parseval's equality which yields

$$\|u - \sum_{\nu \in \Lambda} c_\nu L_\nu\|_{L^2(U,V,d\rho)} = \left(\sum_{\nu \notin \Lambda} \|c_\nu\|_V^2\right)^{\frac{1}{2}},$$

and therefore

$$\|u - u_\Lambda\|_{L^2(U,V,d\rho)} \leq C_1 \left(\sum_{\nu \notin \Lambda} \|c_\nu\|_V^2\right)^{\frac{1}{2}}. \tag{5.11}$$

We also reach (5.11) up to a change in the constant C_1 if $d\rho = w d\mu$ with $w \in L^\infty(U)$.

- **Case 2:** in the case of general ρ, we can still write

$$\|u - \sum_{\nu \in \Lambda} c_\nu L_\nu\|_{L^2(U,V,d\rho)} \leq \|u - \sum_{\nu \in \Lambda} c_\nu L_\nu\|_{L^\infty(U,V)},$$

so that by triangle inequality, we obtain

$$\|u - u_\Lambda\|_{L^2(U,V,d\rho)} \leq C_1 \sum_{\nu \notin \Lambda} \|c_\nu\|_V \|L_\nu\|_{L^\infty(U)}. \tag{5.12}$$

15

The estimates (5.11) and (5.12) suggest to choose for Λ the sets of indices ν corresponding respectively to *the N largest values* of $\|c_\nu\|_V$ and $\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)}$. Of course the exact value of $\|c_\nu\|_V$ is unknown, and therefore a more reasonable objective is to build Λ based on a-priori bounds for $\|c_\nu\|_V$. We shall derive such bounds in the next section. The process of approximating a sequence by retaining its N largest terms is a simple instance of *nonlinear approximation* (see [7] for a general survey) known as *best N-term approximation*. The rate of convergence of this process is well understood, thanks to the following result of Stechkin whose proof is elementary (e.g. [7]).

Lemma 5.5 *Let $0 < p \leq q$ and $\alpha = (\alpha_\nu)_{\nu \in \mathcal{F}}$ be a sequence in $\ell^p(\mathcal{F})$. If $\mathcal{F}_N$ is the set of indices corresponding to the N largest values of $|\alpha_\nu|$, we have*

$$\left(\sum_{\nu \notin \mathcal{F}_N} |\alpha_\nu|^q \right)^{\frac{1}{q}} \leq \|\alpha\|_{\ell^p(\mathcal{F})} N^{-r},$$

where $r := \frac{1}{p} - \frac{1}{q} \geq 0$.

The a-priori bounds that we shall obtain in the next section will also be used to analyze the summability of the sequence $(\|c_\nu\|_V)$ in ℓ^2 and of $(\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)})$ in ℓ^1, and therefore derive the rate of convergence $r > 0$ in (5.4) based on the estimates (5.11)-(5.12) combined with the above lemma.

6 The decay of the Legendre coefficients of u

The decay of the Legendre coefficients of a function depends on its smoothness. For example, one simple way to relate c_ν to $\partial_y^\nu u$ is through Rodrigues' formula which reads

$$L_n(t) = \frac{(-1)^n \sqrt{2n+1}}{2^n n!} \left(\frac{d}{dt} \right)^n ((1 - t^2)^n),$$

when the Legendre polynomials are normalized according to (3.4). For a function $f(t)$ of one variable which is n-times continuously differentiable, we can apply n integrations by parts and obtain a bound for the coefficient $c_n := \int_{-1}^{1} f(t) L_n(t) \frac{dt}{2}$:

$$|c_n| = \frac{\sqrt{2n+1}}{2^n n!} \left| \int_{-1}^{1} (1 - t^2)^n f^{(n)} \frac{dt}{2} \right| \leq \frac{I_n}{2^n n!} \|f^{(n)}\|_{L^\infty([-1,1])},$$

with

$$I_n := \sqrt{2n+1} \int_{-1}^{1} (1 - t^2)^n \frac{dt}{2} = \sqrt{2n+1} \prod_{k=1}^{n} \frac{2k}{2k+1} \quad \text{if } n \geq 1, \ I_0 := 1.$$

The sequence I_n is uniformly bounded. However, it will be sufficient for us to employ the crude bound $I_n \leq I_1^n$ with $I_1 = \frac{2}{3}\sqrt{3} \approx 1.155$, and therefore obtain

$$|c_n| \leq \frac{\beta^n}{n!} \|f^{(n)}\|_{L^\infty([-1,1])},$$

16

with

$$\beta := I_1/2 = 1/\sqrt{3} \approx 0.577. \tag{6.1}$$

This fixes β for the remainder of this paper. Applying similar arguments to $u(x,y)$ in each variable y_j yields

$$\|c_\nu\|_V \leq \frac{\beta^{|\nu|}}{\nu!}\|\partial_y^\nu u\|_{L^\infty(U,V)}. \tag{6.2}$$

We estimate the quantities $\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)}$ in a similar way, replacing I_n by $J_n = \sqrt{2n+1}I_n$ and using the crude bound $J_n \leq 2^n$. This leads to

$$\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)} \leq \frac{1}{\nu!}\|\partial_y^\nu u\|_{L^\infty(U,V)}. \tag{6.3}$$

Combining (4.9) with (6.2) and (6.3), we obtain the following.

Corollary 6.1 *Let $b = (b_j)_{j\geq 1}$ and $d = (d_j)_{j\geq 1}$ be defined by $b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}$ and $d_j = \beta b_j$. We then have with B as in (1.4) for all $\nu \in \mathcal{F}$*

$$\|c_\nu\|_V \leq B\frac{|\nu|!}{\nu!}d^\nu \tag{6.4}$$

and

$$\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)} \leq B\frac{|\nu|!}{\nu!}b^\nu. \tag{6.5}$$

7 Sequence approximation

As explained at the end of §5, the rate of convergence N^{-r} of the spectral approach based on the optimal choice of Λ is related to the properties of ℓ^p summability of the multi-indexed sequences $(\|c_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)})_{\nu\in\mathcal{F}}$. In view of the estimates obtained above for $\|c_\nu\|_V$ and $\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)}$, we need to study the ℓ^p summability of multi-indexed sequences which have the general form

$$\left(\frac{|\nu|!}{\nu!}\alpha^\nu\right)_{\nu\in\mathcal{F}} \tag{7.1}$$

where $\alpha = (\alpha_j)_{j\geq 1}$ is a sequence of positive numbers. Since the rate is either given by $r = \frac{1}{p} - \frac{1}{2}$ or $r = \frac{1}{p} - 1$ (depending on the relation between the measure ρ and the uniform measure μ) and since we are interested in understanding under which circumstances r may be larger than $\frac{1}{2}$, we need to consider values of p smaller than 1.

In this section, we establish simple necessary and sufficient conditions on a sequence α for the ℓ^p summability of the multi-indexed sequence (7.1). Our first result deals with sequences which have the simpler form $(\alpha^\nu)_{\nu\in\mathcal{F}}$.

Lemma 7.1 *For $p \leq 1$, the sequence $(\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $\alpha \in \ell^p(\mathbb{N})$ and $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$. Under these conditions, we have*

$$\|(\alpha^\nu)\|_{\ell^p(\mathcal{F})} \leq \exp\left\{\frac{\|\alpha\|_{\ell^p(\mathbb{N})}^p}{p(1 - \|\alpha\|_{\ell^\infty(\mathbb{N})}^p)}\right\}. \tag{7.2}$$

17

Proof: Assume first that $(\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$. By remarking that for $\nu = e_j$ the Kroeneker sequence $\alpha^\nu = \alpha_j$, we find that $\alpha \in \ell^p(\mathbb{N})$ with $\|\alpha\|_{\ell^p(\mathbb{N})} \leq \|(\alpha^\nu)\|_{\ell^p(\mathcal{F})}$. For each fixed j, the sequence $(\alpha^\nu)_{\nu\in\mathcal{F}}$ contains $(\alpha_j^n)_{n\geq 0}$ as a subsequence corresponding to the indices $\nu = ne_j$ and hence we must have $\alpha_j < 1$. From the fact that $\alpha \in \ell^p(\mathbb{N})$, we see that $\alpha_j \to 0$ as $j \to +\infty$ and hence $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$.

Conversely, if $\alpha \in \ell^p(\mathbb{N})$ and $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$, we can write

$$\|(\alpha^\nu)\|_{\ell^p(\mathcal{F})}^p = \sum_{\nu\in\mathcal{F}}(\alpha^\nu)^p = \sum_{\nu\in\mathcal{F}}\prod_{j\geq 1}\alpha_j^{p\nu_j} = \prod_{j\geq 1}\sum_{n\geq 0}\alpha_j^{np} = \prod_{j\geq 1}\frac{1}{1-\alpha_j^p},$$

where we have used that $\alpha_j < 1$ for all $j \geq 1$. In order to prove the convergence of the infinite product, we remark that

$$\frac{1}{1-\alpha_j^p} \leq 1 + \kappa\alpha_j^p, \quad j = 1,2,\dots \text{ where } \kappa := \frac{1}{1-\|\alpha\|_{\ell^\infty(\mathbb{N})}^p}$$

so that we can write

$$\log\left(\prod_{j\geq 1}\frac{1}{1-\alpha_j^p}\right) \leq \sum_{j\geq 1}\log(1+\kappa\alpha_j^p) \leq \kappa\sum_{j\geq 1}\alpha_j^p = \frac{\|\alpha\|_{\ell^p(\mathbb{N})}^p}{1-\|\alpha\|_{\ell^\infty(\mathbb{N})}^p}$$

which implies the bound (7.2). $\diamond$

We next turn to multi-indexed sequences which have the form (7.1).

Theorem 7.2 *For $p \leq 1$, the sequence $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $\|\alpha\|_{\ell^1(\mathbb{N})} < 1$ and $\alpha \in \ell^p(\mathbb{N})$. One has the estimate*

$$\|(\frac{|\nu|!}{\nu!}\alpha^\nu)\|_{\ell^p(\mathcal{F})} \leq \frac{2}{\eta}\exp\left(\frac{2(1-p)(J(\eta)+\|\alpha\|_{\ell^p(\mathbb{N})}^p)}{p^2\eta}\right), \tag{7.3}$$

where $\eta := (1 - \|\alpha\|_{\ell^1(\mathbb{N})})/2$ and $J(\eta)$ is the smallest positive integer such that $\sum_{j>J}|\alpha_j|^p \leq \frac{\eta}{2}$.

Proof: First notice that $\|(\frac{|\nu|!}{\nu!}\alpha^\nu)\|_{\ell^p(\mathcal{F})}$ does not change if we rearrange the entries in α and therefore we may without loss of generality assume that the nonnegative sequence α is decreasing. For $p = 1$, we have

$$\|(\frac{|\nu|!}{\nu!}\alpha^\nu)\|_{\ell^1(\mathcal{F})} = \sum_{k=0}^\infty(\sum_{j=1}^\infty\alpha_j)^k = \frac{1}{1-\|\alpha\|_{\ell^1(\mathbb{N})}}, \tag{7.4}$$

which gives the theorem in this case. So we consider further only the case $p < 1$.

Assuming that $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$, we notice that the sequence $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ contains $\alpha = (\alpha_j)_{j\geq 1}$ as subsequence corresponding to the indices $\nu = e_j$, and therefore the ℓ^p summability of α is necessary. On the other hand, since $p \leq 1$, the sequence $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ only if it is summable, i.e. it belongs to $\ell^1(\mathcal{F})$. Hence, (7.4) gives that $\|\alpha\|_{\ell^1} < 1$ is necessary.

Conversely, let α be a sequence such that $\alpha \in \ell^p(\mathbb{N})$ and $\|\alpha\|_{\ell^1(\mathbb{N})} < 1$, we shall construct a factorization of α as $\alpha_j = \gamma_j\delta_j$, $j = 0,1,\dots$, with sequences γ and δ satisfying

$$\|\gamma\|_{\ell^1(\mathbb{N})} < 1, \qquad \|\delta\|_{\ell^\infty(\mathbb{N})} < 1, \qquad \|\delta\|_{\ell^{p'}(\mathbb{N})} < \infty, \qquad p' := p/(1-p) > 0. \tag{7.5}$$

18

Having such a factorization of α at hand, we estimate

$$
\begin{aligned}
\sum_{\nu \in \mathcal{F}} \left(\frac{|\nu|!}{\nu!} \alpha^\nu \right)^p &= \sum_{\nu \in \mathcal{F}} \left(\frac{|\nu|!}{\nu!} \gamma^\nu \right)^p \delta^{p\nu} \\
&\leq \left(\sum_{\nu \in \mathcal{F}} \frac{|\nu|!}{\nu!} \gamma^\nu \right)^p \left(\sum_{\nu \in \mathcal{F}} \delta^{\frac{p}{1-p}\nu} \right)^{1-p} = \left\| \left(\frac{|\nu|!}{\nu!} \gamma^\nu \right) \right\|_{\ell^1(\mathcal{F})}^p \| (\delta^\nu) \|_{\ell^{p'}(\mathcal{F})}^p \\
&\leq \ [1 - \|\gamma\|_{\ell^1(\mathbb{N})}]^{-p} \exp \left(\frac{(1-p) \|\delta\|_{\ell^{p'}(\mathbb{N})}^{p'}}{1 - \|\delta\|_{\ell^\infty(\mathbb{N})}^{p'}} \right)
\end{aligned}
\tag{7.6}
$$

where we have first used Hölder's inequality and then we employed (7.4) on the first term and Lemma 7.1 (with p' in place of p and δ in place of α) on the second.

It remains to construct the factor sequences δ and γ satisfying (7.5). To this end, we observe that for every $\eta > 0$, there exists $J(\eta)$ such that

$$
\sum_{j > J(\eta)} |\alpha_j|^p \leq \frac{\eta}{2}.
$$

Choose

$$
\eta := (1 - \|\alpha\|_{\ell^1(\mathbb{N})})/2.
$$

Then $0 < \eta < 1/2$ and we define the factor sequences γ and δ by

$$
\gamma_j := (1+\eta)\alpha_j \ \text{ and } \ \delta_j := \frac{1}{1+\eta} \qquad j \leq J(\eta),
\tag{7.7}
$$

and

$$
\gamma_j = \alpha_j^p \ \text{ and } \ \delta_j = \alpha_j^{1-p}, \qquad j > J(\eta).
\tag{7.8}
$$

Then we read off (7.7) and (7.8) that

$$
\|\delta\|_{\ell^\infty(\mathbb{N})} \leq \max\{(1+\eta)^{-1}, \|\alpha\|_{\ell^\infty(\mathbb{N})}^{1-p}\} \leq \max\{(1+\eta)^{-1}, \|\alpha\|_{\ell^1(\mathbb{N})}^{1-p}\} < 1.
\tag{7.9}
$$

We also have

$$
\|\gamma\|_{\ell^1(\mathbb{N})} \leq (1+\eta)\|\alpha\|_{\ell^1(\mathbb{N})} + \sum_{j > J(\eta)} |\alpha_j|^p \leq (1+\eta)(1-2\eta) + \frac{\eta}{2} \leq 1 - \frac{\eta}{2} < 1.
\tag{7.10}
$$

Finally, we obtain with $p' := p/(1-p)$:

$$
\|\delta\|_{\ell^{p'}(\mathbb{N})}^{p'} \leq J(\eta)(1+\eta)^{-p'} + \|\alpha\|_{\ell^p(\mathbb{N})}^p < \infty.
\tag{7.11}
$$

In order to obtain the bound (7.3), we use (7.6) which states that

$$
\left\| \left(\frac{|\nu|!}{\nu!} \alpha^\nu \right) \right\|_{\ell^p(\mathcal{F})} \leq [1 - \|\gamma\|_{\ell^1(\mathbb{N})}]^{-1} \exp \left(\frac{(1-p) \|\delta\|_{\ell^{p'}(\mathbb{N})}^{p'}}{p(1 - \|\delta\|_{\ell^\infty(\mathbb{N})}^{p'})} \right),
\tag{7.12}
$$

From (7.10), we find that

$$
[1 - \|\gamma\|_{\ell^1(\mathbb{N})}]^{-1} \leq \frac{2}{\eta}.
\tag{7.13}
$$

From (7.9), we infer

$$\|\delta\|_{\ell^\infty(\mathbb{N})} \le \max\{(1+\eta)^{-1}, (1-2\eta)^{1-p}\} \le \max\{1-\frac{\eta}{2}, 1-2(1-p)\eta\} \le 1-(1-p)\frac{\eta}{2},$$

where we have used the fact that $\eta \le \frac{1}{2}$ and therefore

$$1 - \|\delta\|_{\ell^\infty(\mathbb{N})}^{p'} \ge p'(1-p)\frac{\eta}{2} = \frac{p}{2}\eta. \tag{7.14}$$

From (7.11), we infer

$$\|\delta\|_{\ell^{p'}(\mathbb{N})}^{p'} \le J(\eta)(1+\eta)^{-p'} + \|\alpha\|_{\ell^p(\mathbb{N})}^{p} \le J(\eta) + \|\alpha\|_{\ell^p(\mathbb{N})}^{p}. \tag{7.15}$$

Inserting the estimates (7.13), (7.15) and (7.14) inside (7.12) we obtain (7.3). $\diamond$

We now combine the above result with the estimates (6.4) and (6.5), taking as α the sequences b and d. Note that according to (2.4), these sequences respectively satisfy $\|b\|_{\ell^1} \le \kappa$ and $\|d\|_{\ell^1} \le \beta\kappa$ under Assumption 2. We thus obtain the following.

Corollary 7.3 *Assume that* Assumption 3 *holds for some $p \le 1$.*
(i) *If moreover* Assumption 2 *holds with $\kappa := \frac{1}{\beta}$, then $(\|c_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$*
(ii) *If moreover* Assumption 2 *holds with $\kappa := 1$, then $(\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)})_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$.*

Combining Corollary 7.3 with (5.11) and (5.12) and using Lemma 5.5, we obtain the following error estimates between u and u_Λ.

Corollary 7.4 *Assume that* Assumption 3 *holds for some $p \le 1$.*
(i) *If $d\rho = wd\mu$ with $w \in L^\infty(U)$ and if* Assumption 2 *holds with $\kappa := \frac{1}{\beta}$, then there exists a sequence $(\Lambda_N)_{N\in\mathbb{N}} \subset \mathcal{F}$ of index sets Λ of cardinality $N = 1, 2, \ldots$ such that*

$$\|u - u_{\Lambda_N}\|_{L^2(U,V,d\rho)} \le CN^{-r}, \quad r = \frac{1}{p} - \frac{1}{2}. \tag{7.16}$$

(ii) *For a general ρ, if* Assumption 2 *holds with $\kappa := 1$, then there exists a sequence $(\Lambda_N)_{N\in\mathbb{N}} \subset \mathcal{F}$ of index sets Λ_N of cardinality $N = 1, 2, \ldots$ such that*

$$\|u - u_{\Lambda_N}\|_{L^2(U,V,d\rho)} \le CN^{-r}, \quad r = \frac{1}{p} - 1. \tag{7.17}$$

Remark 7.5 *In the case when $\rho = wd\mu$ with $w \in L^\infty$, we always have $r \ge \frac{1}{2}$ and therefore the MC rate (1.7) is outperformed as soon as the ψ_j satisfy the required summability condition.*

Remark 7.6 *In the case where $(\psi_j)_{j\ge1}$ is the Karhúnen-Loève expansion, decay estimates on $\|\psi_j\|_{L^\infty}$ ensuring its ℓ^p-summability are available. These estimates depend on the smoothness properties of the covariance function $C_a(x,y)$, see [20, 18].*

Remark 7.7 *It is obviously interesting to estimate the size of the constants C in the error bounds (7.16) and (7.17). The bound (7.3) on $\left\|\left(\frac{|\nu|!}{\nu!}\alpha^\nu\right)\right\|_{\ell^p(\mathcal{F})}$ obtained in Theorem 7.2 allows us, via Lemma 5.5, to estimate C in (7.16) and (7.17) in terms of the summability properties of the Karhúnen-Loève expansion (1.8) of the input data. However, this bound is not easily computable since it involves the quantity $J(\eta)$ which might actually be arbitrarily large under no assumption other than $\alpha \in \ell^p(\mathbb{N})$. More can be said under the (slightly) stronger assumption that $\alpha \in \ell^q(\mathbb{N})$ for some $q < p$. Assuming without loss of generality that α is non-increasing, we find from Lemma 5.5 that*

$$\sum_{j>J} |\alpha_j|^p \leq \|\alpha\|_{\ell^q(\mathbb{N})}^p J^{\frac{q-p}{q}}.$$

We therefore find that

$$J(\eta) \leq \left(\frac{\eta}{2\|\alpha\|_{\ell^q(\mathbb{N})}^p}\right)^{\frac{q}{q-p}},$$

which leads to the computable bound

$$\|\left(\frac{|\nu|!}{\nu!}\alpha^\nu\right)\|_{\ell^p(\mathcal{F})} \leq \frac{2}{\eta}\exp\left(\frac{4(1-p)\left(\left(\frac{\eta}{2\|\alpha\|_{\ell^q(\mathbb{N})}^p}\right)^{\frac{q}{q-p}} + \|\alpha\|_{\ell^p(\mathbb{N})}^p\right)}{p^2\eta}\right). \tag{7.18}$$

8 Space discretization

Up to this stage, our results allow us to draw a comparison between the convergence rate of MC and deterministic methods in terms of the number N of deterministic unknown functions which need to be determined in such methods. The actual computation of these unknown functions involves space discretization, which is the source of additional approximation error.

The purpose of this last section is to analyze these aspects in order to draw a more exact comparison between the convergence rate of the two methods, now expressed in terms of the total number of degrees of freedom N_{dof}. For the sake of simplicity, we shall focus on space discretizations by the finite element method, although our discussion can be extended to other types of discretization.

In order to establish convergence rates in the above sense, we need to give regularity estimates of the solution u in the physical domain. While this is classical for linear, elliptic equations, we require a-priori estimates *uniform in the parameters $y \in U$*.

For this purpose, additional assumptions are needed. We first recall that when the domain D is either a smooth domain or a convex polyhedron with straight faces, the solution v to the Laplace equation

$$-\Delta v = f \quad \text{in} \quad D, \qquad v|_{\partial D} = 0, \tag{8.1}$$

with $f \in L^2(D)$ belongs to $H_0^1(D) \cap H^2(D)$. This is well-known to yield a convergence rate for the finite element method on families of shape-regular, quasiuniform meshes of meshwidth h: if $(V_h)_{h>0}$ is a one parameter family of finite element spaces associated to a family of shape-regular and quasiuniform partitions of D into simplices of meshwidth $h > 0$, we have the standard approximation estimate

$$\inf_{v_h \in V_h} \|v - v_h\|_V \leq Ch\|v\|_{H^2(D)}, \tag{8.2}$$

i.e. convergence rate $\mathcal{O}(M^{-\frac{1}{d}})$ with $M := \dim(V_h) \sim h^{-d}$. The H^2-smoothness estimate is lost when working on non-convex polyhedrons, however it is well known that the rate $M^{-\frac{1}{d}}$ may sometimes be retained by using continuous, piecewise linear finite elements on certain nonuniform meshes.

This is in particular the case on 2-d polygonal domains with reentrant corners. In order to include this case in our analysis, we introduce the following general assumption.

Assumption 4. *The domain D is such that the subspace*

$$W := \{v \in V \; ; \; \Delta v \in L^2(D)\},$$

equipped with the norm $\|v\|_W := \|\Delta v\|_{L^2}$ has the approximation property

$$\inf_{v_M \in V_M} \|v - v_M\|_V \leq C_a M^{-s} \|v\|_W, \tag{8.3}$$

for some $s > 0$, where $(V_M)_{M>0}$ is a family of finite element spaces such that $\dim(V_M) \leq M$.

Under Assumption 4, the finite element approximation u_M of the solution u of (8.1) satisfies

$$\|u - u_M\|_V \leq C_a M^{-s} \|f\|_{L^2(D)} \tag{8.4}$$

When D is smooth or convex, the space W coincides with $H^2(D) \cap H_0^1(D)$, and $s = \frac{1}{d}$ when V_M is chosen as space of continuous, piecewise linear finite elements on a family of regular, quasiuniform simplicial meshes.

In order to establish similar approximation estimates on the solution of the problem (1.1) with spatially inhomogeneous random coefficients, we need a smoothness assumption on these coefficients.

Assumption 5. *There exists a constant $C_r > 0$ such that*

$$\|\nabla a\|_{L^\infty(U, L^\infty(D))} := \sup_{y \in U} \|\nabla a(\cdot, y)\|_{L^\infty(D)} \leq C_r a_{min}. \tag{8.5}$$

Since (1.1) can be rewritten

$$-\Delta u = \frac{1}{a}[f + \nabla a \cdot \nabla u] =: g,$$

we can estimate

$$\|g\|_{L^2} \leq \frac{1}{a_{\min}}[\|f\|_{L^2(D)} + C_r \|f\|_{V^*}] \leq \frac{1 + C_r C_P}{a_{\min}} \|f\|_{L^2(D)}, \tag{8.6}$$

where C_P denotes the Poincaré constant of D. We thus obtain from Assumption 5 a smoothness estimate for the solution of (1.1):

$$\|u\|_{L^\infty(U,W)} \leq C_2 := \frac{1 + C_r C_P}{a_{\min}} \|f\|_{L^2(D)}. \tag{8.7}$$

22

For comparison purposes, we now establish a convergence estimate for the MC methods with space discretization. Let M be fixed and for each instance a_i, let $u_{i,M} \in V_M$ be the Galerkin projection of u_i onto V_M which is defined by

$$\text{find} \quad u_{i,M} \in V_M : \quad \int_D a_i \nabla u_{i,M} \nabla v_M = \int_D f v_M \quad \text{for all } v_M \in V_M.$$

We define the corresponding approximation to the mean field by

$$\overline{u}_{N,M} := \frac{1}{N} \sum_{i=1}^{N} u_{i,M}.$$

Combining our assumptions with Cea's lemma yields

$$\|u_i - u_{i,M}\|_V \leq C_1 \inf_{v_M \in V_M} \|u_i - v_M\|_V \leq C_a C_1 M^{-s} \|u_i\|_W \leq C_a C_1 C_2 M^{-s},$$

for each $i = 1, \cdots, N$, with $C_1 := \sqrt{\frac{a_{\max}}{a_{\min}}}$. We therefore have

$$\|\overline{u}_N - \overline{u}_{N,M}\|_V \leq \frac{1}{N} \sum_{i=1}^{N} \|u_i - u_{i,M}\|_V \leq C_a C_1 C_2 M^{-s}.$$

Combining this with (1.7), we obtain

$$\mathbb{E}(\|\overline{u} - \overline{u}_{N,M}\|_V) \leq B N^{-\frac{1}{2}} + C_a C_1 C_2 M^{-s}. \tag{8.8}$$

The total number of degrees of freedom appearing in Monte-Carlo Finite Element simulation with N "samples" is $N_{dof} := NM$. To optimize estimate (8.8) with respect to a given total number of degrees of freedom $N_{dof} := NM$, we take $N \sim M^{2s}$. Then $N_{dof} \sim M^{2s+1}$ and we obtain the error estimate

$$\mathbb{E}(\|\overline{u} - \overline{u}_{N,M}\|_V) \leq C N_{dof}^{-\frac{s}{2s+1}}, \tag{8.9}$$

where the constant C depends on B, C_a, C_1 and C_2.

We next turn to the deterministic method. We incorporate the space discretization as follows: for any subset $\Lambda \subset \mathcal{F}$ of finite cardinality and any vector $\mathcal{M} = (M_\nu)_{\nu \in \Lambda}$ of positive integers, we define the approximation space

$$X_{\Lambda,\mathcal{M}} := \{ v_{\Lambda,\mathcal{M}}(x,y) = \sum_{\nu \in \Lambda} v_\nu(x) L_\nu(y) \; ; \; v_\nu \in V_{M_\nu} \}.$$

We define the corresponding Galerkin approximation $u_{\Lambda,\mathcal{M}} = \sum_{\nu \in \Lambda} u_{\nu,\mathcal{M}} L_\nu \in X_{\Lambda,M}$ to u as the unique solution to

$$B(u_{\Lambda,\mathcal{M}}, v_{\Lambda,\mathcal{M}}) = F(v_{\Lambda,\mathcal{M}}), \tag{8.10}$$

for all $v_{\Lambda,\mathcal{M}} \in X_{\Lambda,\mathcal{M}}$, where B and F are defined by (5.2). The total number of degrees of freedom is now given by

$$N_{dof} = \sum_{\nu \in \Lambda} M_\nu.$$

We first mimic the analysis of the Galerkin approximation in §3: from Cea's lemma, we get

$$\|u - u_{\Lambda,\mathcal{M}}\|_{L^2(U,V,d\rho)} \leq C_1 \|u - \sum_{\nu \in \Lambda} c_{\nu,M_\nu} L_\nu\|_{L^2(U,V,d\rho)}, \tag{8.11}$$

for any $c_{\nu,M_\nu} \in V_{M_\nu}$. Specifically, we take c_{ν,M_ν} to be the V-orthogonal projection of the Legendre coefficient c_ν onto V_{M_μ}. Similar to the discussion in §3, we distinguish two cases:

- **Case 1:** if $d\rho = d\mu$, we obtain by orthogonality

$$\|u - u_{\Lambda,\mathcal{M}}\|_{L^2(U,V,d\rho)} \leq C_1 \left(\sum_{\nu \notin \Lambda} \|c_\nu\|_V^2 + \sum_{\nu \in \Lambda} \|c_\nu - c_{\nu,M_\nu}\|_V^2 \right)^{\frac{1}{2}}. \tag{8.12}$$

We also reach (8.12) up to a change in the constant C_1 if $d\rho = w d\mu$ with $w \in L^\infty(U)$.

- **Case 2:** in the case of general ρ, we obtain by the triangle inequality

$$\|u - u_{\Lambda,\mathcal{M}}\|_{L^2(U,V,d\rho)} \leq C_1 \left(\sum_{\nu \notin \Lambda} \|c_\nu\|_V \|L_\nu\|_{L^\infty(U)} + \sum_{\nu \in \Lambda} \|c_\nu - c_{\nu,M_\nu}\|_V \|L_\nu\|_{L^\infty(U)} \right). \tag{8.13}$$

The right hand side of the estimates (8.12) and (8.13) are similar to (5.11) and (5.12) up to an additional term reflecting space discretization. From (8.3), we obtain

$$\|c_\nu - c_{\nu,M_\nu}\|_V \leq C_a M_\nu^{-s} \|c_\nu\|_W.$$

Under the assumptions of Corollary 7.4, we thus obtain from (8.12) in the first case

$$\|u - u_{\Lambda,\mathcal{M}}\|_{L^2(U,V,d\rho)} \leq C \left(N^{-2r} + \sum_{\nu \in \Lambda} M_\nu^{-2s} \|c_\nu\|_W^2 \right)^{\frac{1}{2}}, \tag{8.14}$$

where $N := \#(\Lambda)$ and from (8.13) in the second case

$$\|u - u_{\Lambda,\mathcal{M}}\|_{L^2(U,V,d\rho)} \leq C \left(N^{-r} + \sum_{\nu \in \Lambda} M_\nu^{-s} \|c_\nu\|_W \|L_\nu\|_{L^\infty(U)} \right). \tag{8.15}$$

Based on these estimates, we optimize the discretization parameter $\mathcal{M} = (M_\nu)_{\nu \in \Lambda}$ in order to estimate the best possible convergence rate for the deterministic method in terms of the total number of degrees of freedom. The optimization problem to be solved consists in minimizing the number of degrees of freedom under the constraint that the additional term reflecting space discretization remains of the same order as the first term reflecting discretization in the y variable, i.e.

$$\text{Min}\{\sum_{\nu \in \Lambda} M_\nu : \sum_{\nu \in \Lambda} M_\nu^{-2s} \|c_\nu\|_W^2 \leq N^{-2r}\}, \tag{8.16}$$

in the first case and

$$\text{Min}\{\sum_{\nu \in \Lambda} M_\nu : \sum_{\nu \in \Lambda} M_\nu^{-s} \|c_\nu\|_W \|L_\nu\|_{L^\infty(U)} \leq N^{-r}\}, \tag{8.17}$$

in the second case. We solve both problems by treating the M_ν as continuous variables, up to finally taking the integer value of the solution. For (8.16), introducing a Lagrange multiplier, we obtain

$$M_\nu = A^{\frac{1}{1+2s}} \|c_\nu\|_W^{\frac{2}{1+2s}} \qquad \forall \nu \in \Lambda$$

where the value of A is given by

$$N^{-2r} = \sum_{\nu \in \Lambda} M_\nu^{-2s} \|c_\nu\|_W^2 = A^{-1} \sum_{\nu \in \Lambda} M_\nu = A^{-\frac{2s}{1+2s}} \sum_{\nu \in \Lambda} \|c_\nu\|_W^{\frac{2}{1+2s}}.$$

Two situations may occur depending on the summability properties of the sequence $(\|c_\nu\|_W)_{\nu \in \mathcal{F}}$:

- If $(\|c_\nu\|_W)_{\nu \in \mathcal{F}} \in \ell^p(\mathcal{F})$ with $p = \frac{2}{1+2s}$, we obtain that $A^{-\frac{2s}{1+2s}} \sim N^{-2r}$. It follows that

$$N_{dof} = \sum_{\nu \in \Lambda} M_\nu = AN^{-2r} \sim A^{\frac{1}{1+2s}} \sim N^{\frac{r}{s}}.$$

We therefore obtain the convergence rate

$$\|u - u_{\Lambda,\mathcal{M}}\|_{L^2(U,V,d\rho)} \le CN_{dof}^{-s}. \tag{8.18}$$

- If $(\|c_\nu\|_W)_{\nu \in \mathcal{F}} \in \ell^p(\mathcal{F})$ for some $p > \frac{2}{1+2s}$, we can estimate A by using Hölder's inequality as follows:

$$A^{\frac{2s}{1+2s}} N^{-2r} = \sum_{\nu \in \Lambda} \|c_\nu\|_W^{\frac{2}{1+2s}} \le \left(\sum_{\nu \in \Lambda} \|c_\nu\|_W^p \right)^{\frac{2}{p+2sp}} N^{1 - \frac{2}{p+2sp}} = CN^\delta,$$

with $\delta := 1 - \frac{2}{p+2sp} > 0$. This leads to

$$N_{dof} = \sum_{\nu \in \Lambda} M_\nu = AN^{-2r} \sim N^{\frac{(2r+\delta)(1+2s)}{2s} - 2r} = N^{\frac{2r+\delta(1+2s)}{2s}}.$$

We therefore obtain the convergence rate

$$\|u - u_{\Lambda,\mathcal{M}}\|_{L^2(U,V,d\rho)} \le CN_{dof}^{-\frac{2sr}{2r+\delta(1+2s)}} \tag{8.19}$$

Remark 8.1 *The first estimate* (8.18) *shows that if the sequence* $(\|c_\nu\|_W)_{\nu \in \mathcal{F}}$ *is sufficiently concentrated, the rate of convergence of our method is similar to solving one single deterministic problem and therefore optimally fast. On the other hand, since we have by Parseval's equality*

$$\sum_{\nu \in \mathcal{F}} \|c_\nu\|_W^2 = \|u\|_{L^2(U,W,d\mu)}^2 \le \|u\|_{L^\infty(U,W)}^2 \le C_2^2$$

with C_2 *as in* (8.7), *we are always ensured that* $(\|c_\nu\|_W)_{\nu \in \mathcal{F}} \in \ell^2(\mathcal{F})$. *In the worst case* $p = 2$, *the rate of convergence is given by the second estimate* (8.19) *with* $\delta = \frac{2s}{1+2s}$, *therefore* $N^{-\frac{2sr}{2r+2s}}$ *which is still faster than the MC rate* (8.9) *if* $r > \frac{1}{2}$, *since* $\frac{2sr}{2r+2s} - \frac{s}{1+2s} > 0$ *then.*

By applying a similar analysis to the optimization problem (8.17) we obtain that (8.18) holds provided that $(\|c_\nu\|_W \|L_\nu\|_{L^\infty(U)})_{\nu \in \mathcal{F}} \in \ell^p(\mathcal{F})$ with $p = \frac{1}{1+s} < 1$. If this sequence belongs to $\ell^p(\mathcal{F})$ for some $p > \frac{1}{1+s}$, we obtain the final error estimate

$$\|u - u_\Lambda\|_{L^2(U,V,d\rho)} \le CN_{dof}^{-\frac{sr}{r+\delta(1+s)}}, \tag{8.20}$$

with $\delta := 1 - \frac{1}{p+sp} > 0$.

In view of these results, our last task is therefore to analyze the ℓ^p-summability properties of the sequences $(\|c_\nu\|_W)_{\nu \in \mathcal{F}}$ and $(\|c_\nu\|_W \|L_\nu\|_{L^\infty(U)})_{\nu \in \mathcal{F}}$. We proceed in a similar way as for the sequences $(\|c_\nu\|_V)_{\nu \in \mathcal{F}}$ and $(\|c_\nu\|_V \|L_\nu\|_{L^\infty(U)})_{\nu \in \mathcal{F}}$, estimating first the derivatives $\|\partial_y^\nu u\|_{L^\infty(U,W)}$.

Theorem 8.2 *Let the sequence* $b(\varepsilon) = (b_j(\varepsilon))_{j \geq 1}$ *be defined by*

$$b_j(\varepsilon) := b_j + \varepsilon(\|\nabla \psi_j\|_{L^\infty(D)} + C_r\|\psi_j\|_{L^\infty(D)}),$$

where C_r is the constant in Assumption 5, $b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}$, $j = 1, 2, ...$ *and $\varepsilon > 0$ is arbitrary. We then have*

$$\|\partial_y^\nu u\|_{L^\infty(U,W)} \leq C_3 |\nu|! b(\varepsilon)^\nu \tag{8.21}$$

for all $\nu \in \mathcal{F}$, where $C_3 := (B + \|f\|_{L^2(D)})\frac{1+C_rC_P}{\varepsilon a_{\min}}$.

Proof: for fixed $y \in U$, $\nu \in \mathcal{F}$ we introduce the notation $v_\nu(x) := \nabla \cdot (a(x,y)\nabla\partial_y^\nu u(x,y))$, and remark that the function $\partial_y^\nu u(x,y)$ is the solution to the elliptic problem

$$-\nabla \cdot (a(x,y)\nabla\partial_y^\nu u(x,y)) = -v_\nu(x) \quad \text{in} \quad D, \qquad \partial_y^\nu u(x,y)|_{\partial D} = 0. \tag{8.22}$$

Using the regularity estimate (8.7), we obtain that

$$\|\partial_y^\nu u(\cdot, y)\|_W \leq \frac{1+C_rC_P}{a_{\min}}\|v_\nu\|_{L^2(D)}. \tag{8.23}$$

We now estimate $\|v_\nu\|_{L^2(D)}$. To this end, we start from the identity (4.10): for any $y \in U$ and any $v \in V$

$$\int_D \nabla \cdot (a(x,y)\nabla\partial_y^\nu u(x,y))v(x)dx + \sum_{\{j:\, \nu_j \neq 0\}} \nu_j \int_D [\nabla\psi_j(x) \cdot \nabla\partial_y^{\nu-e_j} u(x,y) + \psi_j(x)\Delta\partial_y^{\nu-e_j} u(x,y)]v(x)dx = 0.$$

Taking here $v = v_\nu \in V$ and using the Cauchy-Schwarz inequality we obtain

$$\|v_\nu\|_{L^2(D)} \leq \sum_{\{j:\, \nu_j \neq 0\}} \nu_j \left(\|\nabla\psi_j\|_{L^\infty(D)}\|\partial_y^{\nu-e_j} u(\cdot, y))\|_V + \|\psi_j\|_{L^\infty(D)}\|\Delta\partial_y^{\nu-e_j} u(\cdot, y))\|_{L^2(D)}\right). \tag{8.24}$$

We next observe that it follows from (8.22) and (8.6) with (8.5) in Assumption 5 that for any $y \in U$

$$\|\Delta(\partial_y^{\nu-e_j} u(\cdot, y))\|_{L^2(D)} \leq \frac{1}{a_{\min}}\|v_{\nu-e_j}(\cdot, y)\|_{L^2(D)} + C_r\|\partial_y^{\nu-e_j} u(\cdot, y)\|_V.$$

Inserting this in (8.24) implies

$$\|v_\nu\|_{L^2(D)} \leq \sum_{\{j:\, \nu_j \neq 0\}} \nu_j \left((\|\nabla\psi_j\|_{L^\infty(D)} + C_r\|\psi_j\|_{L^\infty(D)})\|\partial_y^{\nu-e_j} u(\cdot, y))\|_V + b_j\|v_{\nu-e_j}\|_{L^2(D)}\right) \tag{8.25}$$

with b_j as in (4.8). Multiplying (8.25) by $\varepsilon > 0$ and adding it to the estimate (4.11) established in the proof of Theorem 4.3, we obtain

$$\|\partial_y^\nu u(\cdot, y)\|_V + \varepsilon\|v_\nu\|_{L^2(D)} \leq \sum_{\{j:\, \nu_j \neq 0\}} \nu_j b_j(\|\partial_y^{\nu-e_j} u(\cdot, y)\|_V + \varepsilon\|v_{\nu-e_j}\|_{L^2(D)})$$
$$+ \sum_{\{j:\, \nu_j \neq 0\}} \nu_j \varepsilon(\|\nabla\psi_j\|_{L^\infty(D)} + C_r\|\psi_j\|_{L^\infty(D)})\|\partial_y^{\nu-e_j} u(\cdot, y))\|_V,$$

and therefore

$$\|\partial_y^\nu u(\cdot, y)\|_V + \varepsilon\|v_\nu\|_{L^2(D)} \leq \sum_{\{j:\, \nu_j \neq 0\}} \nu_j b_j(\varepsilon)(\|\partial_y^{\nu-e_j} u(\cdot, y)\|_V + \varepsilon\|v_{\nu-e_j}\|_{L^2(D)}). \tag{8.26}$$

Using the same reasoning by induction as in the proof of Theorem 4.3, we infer from (8.26) that

$$\|\partial_y^\nu u(\cdot,y)\|_V + \varepsilon\|v_\nu\|_{L^2(D)} \leq (\|u(\cdot,y)\|_V + \varepsilon\|\nabla\cdot(a(x,y)\nabla u(x,y))\|_{L^2(D)})|\nu|!b(\varepsilon)^\nu$$
$$\leq (B + \|f\|_{L^2(D)})|\nu|!b(\varepsilon)^\nu$$

for all $\nu \in \mathcal{F}$. We thus have

$$\|v_\nu\|_{L^2(D)} \leq \frac{1}{\varepsilon}(B + \|f\|_{L^2(D)})|\nu|!b(\varepsilon)^\nu$$

which, together with (8.23), concludes the proof.   ◇

We next proceed similar to the study of the summability of $(\|c_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|c_\nu\|_V\|L_\nu\|_{L^\infty(U)})_{\nu\in\mathcal{F}}$. For this purpose, we introduce the following analog to Assumption 3 for the functions $\nabla\psi_j$.

Assumption 6. *The sequence* $(\|\nabla\psi_j\|_{L^\infty(D)})_{j\geq 1}$ *belongs to* $\ell^p(\mathbb{N})$ *for some* $p < 1$:

$$\sum_{j\geq 1}\|\nabla\psi_j\|_{L^\infty(D)}^p < \infty$$

Using the fact that ε can be chosen arbitrarily small in the statement of Theorem 8.2, we reach the following analog to Corollary 7.3.

Corollary 8.3 *Assume that* Assumptions 4 *and* 5 *hold, and that* Assumptions 3 *and* 6 *hold with the same* $p \leq 1$.
(i) *If* Assumption 2 *holds with* $\kappa := \frac{1}{\beta}$, *then* $(\|c_\nu\|_W)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$ *and we obtain the error bound* (8.18) *if* $p = \frac{2}{2s+1}$ *and* (8.19) *if* $p > \frac{2}{2s+1}$.
(ii) *If* Assumption 2 *holds with* $\kappa := 1$, *then* $(\|c_\nu\|_W\|L_\nu\|_{L^\infty(U)})_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$, *and we obtain the error bound* (8.20).

9   Conclusion and perspectives

The deterministic approach which is studied in this paper outperforms the Monte-Carlo approach in terms of convergence rate, provided that the expansion of the random coefficient a in the basis ψ_j has some summability properties in the L^∞ norm. Our analysis is restricted to random coefficients which are uniformly elliptic in the sense of Assumption A1. We expect that similar conclusions hold in different settings, in particular log-normal coefficients, i.e. $a(x,\omega) := \exp(b(x,\omega))$ where b is a Gaussian random field. In this setting, the Legendre polynomials need to be replaced by the Hermite polynomials which are orthonormal with respect to the Gaussian measure. We remark that the analytic regularity results Theorems 4.3 and 8.2 were obtained by real-variable inductive arguments. A different avenue of their proof is through techniques of several complex variables; this is explored in the forthcoming work [6].

This paper has been concerned with establishing results on the approximation of solutions to stochastic and parametric problems by finite dimensional adaptively chosen Galerkin subspaces. Our analysis *does not* propose a specific algorithm for identifying these subspaces. Our results

should therefore rather be considered as a benchmark for the convergence analysis of numerical methods for the approximation of parametric and stochastic PDE's in the x and y variables. The two most commonly used numerical methods are Galerkin projections (as described in §5 and §8) and collocation [2, 12, 13]. In order to retrieve the same convergence rates which are proved in the present paper, such methods need to be developed within an adaptive framework, with the goal of selecting proper sets Λ_N and finite element spaces V_ν throughout the numerical computation. This will be a subject of our future work but we can provide some preliminary comments on finding good Galerkin subspaces.

In the proofs of our convergence theorems, we establish *a-priori* estimates for the Legendre coefficients $\|c_\nu\|_V$. This suggests a first strategy that consists in choosing Λ by selecting the N largest ones from the available a-priori estimates. Another approach that might be more effective is to adaptively build Λ through *a-posteriori* error estimates. This means that we start from the coarse set $\Lambda^0 = \{0\}$ and recursively construct a *nested* sequence Λ^n using error indicators. Such space refinement strategies have been explored in the simpler context of adaptive wavelet approximation of the solution to a single elliptic PDE's, see in particular [5, 16]. In these works, it is shown that a standard bulk chasing strategy based on a-posteriori error indicators leads to optimal convergence rates. The adaptation of this approach to the parametric and stochastic PDE's addressed in this paper is currently under investigation. A critical issue is also the proper adaptive tuning of the space discretization with respect to the different indices ν as revealed by our analysis of §8.

Finally, let us reiterate that an intrinsic weakness in the deterministic approach is that it assumes a complete knowledge on the probability distribution of the coefficients, while Monte-Carlo is applicable when we only have a sample of independent instances at our disposal. In such a case, one may still hope to construct a deterministic solution $u_\Lambda \in V_\Lambda$, either by the collocation method or by a Galerkin system similar to (5.3) in which the integrals over U with respect to the unkown measure $d\rho$ are replaced by computable empirical expectations based on the available samples.

References

[1] I. Babuska, R. Tempone and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal. **42**(2004), 800–825.

[2] I. Babuška, F. Nobile and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Num. Anal., **45**(2007), 1005–1034.

[3] P.G. Ciarlet, *The Finite Element Methods for Elliptic Problems*, Elsevier, Amsterdam 1978.

[4] A. Cohen, *Numerical analysis of wavelet methods*, Elsevier, Amsterdam 2003.

[5] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations - convergence rates*, Math. Comp. **70**(2001), 27-75.

[6] A. Cohen, R. DeVore, and C. Schwab, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's*, to appear in Analysis and Application, 2010.

[7] R. DeVore, *Nonlinear Approximation*, Acta Numerica **7**(1998), 51–150.

[8] R. Ghanem and P. Spanos, *Spectral techniques for stochastic finite elements*, Arch. Comput. Meth. Eng. **4**(1997), 63–100.

[9] G. E. Karniadakis and D. B. Xiu, *The Wiener-Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comp. **24**(2002), 619–644.

[10] M. Kleiber and T. D. Hien, *The stochastic finite element methods*, John Wiley & Sons, Chichester, 1992.

[11] M. Ledoux and M. Talagrand, *Probability in Banachspaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete, vol. 23, Springer Verlag, Berlin, 1991.

[12] F. Nobile, R. Tempone and C.G. Webster, *A sparse grid stochastic collocation method for elliptic partial differential equations with random input data* , SIAM J. Num. Anal. **46**(2008), 2309–2345.

[13] F. Nobile, R. Tempone and C.G. Webster, *An anisotropic sparse grid stochastic collocation method for elliptic partial differential equations with random input data* SIAM J. Num. Anal. **46**(2008), 2411–2442.

[14] T. von Petersdorff and Ch. Schwab, *Sparse Finite Element Methods for Operator Equations with Stochastic Data* Applications of Mathematics **51**(2006) 145–180.

[15] W. Schoutens, *Stochastic processes and orthogonal polynomials*, Lecture notes in statistics vol. 146, Springer-Verlag, New York, 2000.

[16] T. Gantumur, H. Harbecht, and R. Stevenson, *An adaptive wavelet method without coarsening of the iterands*, Math. Comp. **76** (2007), 615-629.

[17] G. Rozza, D.B.P. Huynh, and A.T. Patera, *Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations application to transport and continuum mechanics*, Archives of Computational Methods in Engineering, **15**(2008), 229–275.

[18] Ch. Schwab and R. Todor, *Karhúnen-Loève Approximation of Random Fields by Generalized Fast Multipole Methods*, Journal of Computational Physics **217**(2006), 100–122.

[19] S. A. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions*, Dokl. Akad. Nauk SSSR **4**(1963), 240–243.

[20] R. Todor, *Robust eigenvalue computation for smoothing operators*, SIAM J. Num. Anal. **44**(2006), 865–878.

[21] R. Todor and Ch. Schwab, *Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients*, IMA Journ. Numer. Anal. **45**(2007), 232-261.

[22] N. Wiener, *The homogeneous chaos*, Amer. J. Math **60**(1938), 897–936.

Albert Cohen

UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France

CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France

cohen@ann.jussieu.fr

Ronald DeVore

Department of Mathematics, Texas A& M University, College Station, TX 77843

rdevore@math.tamu.edu

Christoph Schwab

Seminar for Applied Mathematics, ETH Zürich, CH 8092 Zürich

schwab@math.ethz.ch