Computers and Mathematics with Applications 67 (2014) 732–751

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Convergence of quasi-optimal Stochastic Galerkin methods for a class of PDES with random coefficients

Joakim Beck [a], Fabio Nobile [b,c,*], Lorenzo Tamellini [b,c], Raúl Tempone [a]

[a] Applied Mathematics and Computational Science, 4700, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia

[b] MOX - Modellistica e Calcolo Scientifico, Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133, Milano, Italy

[c] CSQI - MATHICSE, Ecole Politechnique Fédérale Lausanne, Station 8, CH 1015, Lausanne, Switzerland

ARTICLE INFO

Keywords:
Uncertainty quantification
Elliptic PDEs with random data
Multivariate polynomial approximation
Best M-terms polynomial approximation
Stochastic Galerkin method
Subexponential convergence

ABSTRACT

In this work we consider quasi-optimal versions of the Stochastic Galerkin method for solving linear elliptic PDEs with stochastic coefficients. In particular, we consider the case of a finite number N of random inputs and an analytic dependence of the solution of the PDE with respect to the parameters in a polydisc of the complex plane $\mathbb{C}^N$. We show that a quasi-optimal approximation is given by a Galerkin projection on a weighted (anisotropic) total degree space and prove a (sub)exponential convergence rate. As a specific application we consider a thermal conduction problem with non-overlapping inclusions of random conductivity. Numerical results show the sharpness of our estimates.

1. Introduction

Partial differential equations with stochastic coefficients have been the subject of growing interest in the scientific community, as they conveniently describe situations in which the coefficients of the PDE are calibrated from noisy and limited measurements and a probabilistic uncertainty model is associated to them. In this context, one may be interested in computing statistics like mean or correlation of the solution of the PDE or possibly statistics of some observables of it, usually called "quantities of interest".

Sampling strategies are widely used to this end, ranging from plain Monte Carlo method to more sophisticated sampling techniques. However, in some cases it is possible to show that the solution is very smooth with respect to the random coefficients, and thus it may be reasonable to use polynomial approximations. In this work, we focus on linear elliptic equations with random diffusion coefficients. These equations exhibit an analytic dependence of the solution on the random input parameters, see e.g. [1–6].

Two relevant polynomial approximation strategies that can be conveniently applied to the problem at hand are the Stochastic Galerkin [1,7–10] and the Stochastic Collocation methods [2,11–13], which are a projection technique and an interpolation technique, respectively. In this work, we reconsider the quasi-optimal Stochastic Galerkin method proposed in the previous work [3], and provide rigorous convergence results in the special case in which the analyticity region contains a polydisc in the complex plane $\mathbb{C}^N$. Observe that in this context "quasi-optimal" means that the proposed methods are optimal with respect to upper bounds of the error, that we observe numerically to be quite sharp.

* Corresponding author at: CSQI - MATHICSE, Ecole Politechnique Fédérale Lausanne, Station 8, CH 1015, Lausanne, Switzerland. Tel.: +41 21 69 35411; fax: +41 21 69 35411.

E-mail addresses: joba@kth.se (J. Beck), fabio.nobile@epfl.ch, fabio.nobile@polimi.it (F. Nobile), lorenzo.tamellini@mail.polimi.it (L. Tamellini), raul.tempone@kaust.edu.sa (R. Tempone).

0898-1221/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.camwa.2013.03.004

In particular, we will derive, under the aforementioned assumptions, the decay of the coefficients of the polynomial expansion of the solution, following the proof in [14] (see also [4]). Next, following the construction of the quasi-optimal polynomial space proposed in [3] (and to some extent also in [4]) we will show that the well-known total degree polynomial space is a quasi-optimal choice for the Stochastic Galerkin method for the class of problems we are considering. We will then derive the corresponding convergence estimates with two different approaches. The first one is based on Taylor expansion and is suitable for isotropic problems; the second one is based on the summability properties of the estimates of the Legendre coefficients of the solution and can be used in an anisotropic setting.

The class of problems that satisfy the analyticity assumption we consider here will be specified in the following (see Remark 3). In particular, it includes the example of a thermal conduction problem with non-overlapping inclusions of random conductivity, originally proposed in [15]. Hence, we will be able to reinterpret the numerical results there obtained in view of the estimates shown here. In particular, it will clearly appear that the theoretical estimates we propose capture correctly the behavior that we observe numerically for the Legendre coefficients and the more than algebraic convergence rate of the global Galerkin error. However, they overestimate considerably the constants in the estimates. Nevertheless, they can be used as the correct ansatz to be fitted by numerical data, resulting in mixed a-priori/a-posteriori methods.

It is worth noting that the analyticity assumption we consider here does not include diffusion coefficients resulting from a truncated Karhunen–Loève expansion of a correlated random field. For such problems, a convergence estimate for the quasi-optimal Stochastic Galerkin method is provided in [5,6], where however no explicit construction of the corresponding polynomial space is given. A possible a-priori formula to this end is given in [3] (the so-called "TD-FC" polynomial approximation).

As an alternative to a-priori constructions, [16,17] propose constructions of quasi-optimal polynomial spaces with adaptive strategies. In particular, the work [17] presents a perturbation method restricted to a small-noise assumption, while [16] presents some algorithms based on Taylor expansions, and tests them also on problems that satisfy the same analyticity assumption considered here. Although very attractive, the main drawback of fully adaptive methods is the cost of exploration of the space of polynomials, that may not be negligible in high dimensions and can be avoided if the correct space of polynomials is prescribed by combining a-priori information with a-posteriori estimates.

The rest of this work is organized as follows: after having detailed in Section 2 the problem at hand and stated Assumption A3 on the analyticity requirements for the solutions considered, we will briefly review the Stochastic Galerkin methodology in Section 3. Section 4 presents then the convergence result for quasi-optimal Stochastic Galerkin method, while Section 5 shows that the solution of a generic "inclusions problem" satisfies the analyticity assumptions. Section 6 will recall the details of the inclusions test presented in [15] and show some numerical results that confirm the sharpness of the proposed estimates. Finally, Section 7 will draw some conclusions and perspectives.

2. Problem setting

2.1. A linear elliptic PDE with stochastic coefficients

Let D be a convex polygonal domain in $\mathbb{R}^d$, and let $(\Omega, \mathcal{F}, \mu)$ be a complete probability space, Ω being the set of outcomes, $\mathcal{F} \subset 2^\Omega$ the σ-algebra of events and $\mu : \mathcal{F} \to [0, 1]$ a probability measure. In this work we focus on the stochastic elliptic problem

Problem 1 (*Strong Formulation*)**.** Find a random field $u : \overline{D} \times \Omega \to \mathbb{R}$, such that μ-almost surely there holds:

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

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

We will work under the following assumptions on the random field $a(\mathbf{x}, \omega)$:

Assumption A1 (*Continuity and Coercivity*)**.** The coefficient $a(\cdot, \omega)$ is a strictly positive and bounded function over D for each random event $\omega \in \Omega$, i.e. there exist two positive constants $\infty > a_{\max} > a_{\min} > 0$ such that $a_{\min} \leq a(\mathbf{x}, \omega) \leq a_{\max}$ μ-almost surely $\forall\, \mathbf{x} \in D$.

Assumption A2 (*"Finite Dimensional Noise Assumption"*)**.** The diffusion coefficient $a(\mathbf{x}, \omega)$ can be parametrized using a vector of N real-valued random variables, namely

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

Such random variables are independent and uniformly distributed, $\mathbf{y}(\omega) = (y_1(\omega), \ldots, y_N(\omega))^T : \Omega \to \Gamma \subset \mathbb{R}^N$, $\Gamma = \Gamma_1 \times \Gamma_2 \times \cdots \times \Gamma_N$. Without loss of generality, we further assume $\Gamma_i = [-1, 1]$, so that the joint probability density function of $\mathbf{y}$, $\varrho : \Gamma \to \mathbb{R}_+$, factorizes as $\varrho(\mathbf{y}) = \prod_{n=1}^N \varrho_n(y_n)$, with $\varrho_n = \frac{1}{2}$.

Assumptions A1 and A2 deserve some comments. First, as an immediate consequence of Assumption A1 and Lax–Milgram's Lemma we have well-posedness of problem (1) for μ-almost every $\omega \in \Omega$.

Next, under Assumption A2 the solution u of (1) depends on the single realization $\omega \in \Omega$ only through the value taken by the random vector $\mathbf{y}$. We can therefore replace the probability space $(\Omega, \mathcal{F}, \mu)$ with $(\Gamma, B(\Gamma), \varrho(\mathbf{y})d\mathbf{y})$, where $B(\Gamma)$ denotes the Borel σ-algebra on Γ and $\varrho(\mathbf{y})d\mathbf{y}$ is the measure of the vector $\mathbf{y}$.

Finally, we observe that more general problems can be addressed within this setting. In particular, problems depending on a set of N non-uniform random variables $z_1, \ldots, z_N$ may be included in this setting by introducing a non-linear map $y_i = \Theta(z_i)$ that transforms each of them into uniform random variables, following the well known theory on copulas, see [18]. In the case a mapping Θ is not available, one could still reduce the problem to the uniform case, by introducing an auxiliary density $\hat{\varrho} = \frac{1}{2^N}$ as suggested in [19]. This will lead to analogous error estimates as those derived in this work, up to a multiplicative constant factor proportional to $\left\|\varrho/\hat{\varrho}\right\|_{L^\infty(\Omega)}$. Even the assumption of independence of the random variables, although very convenient for the development of the tensorized techniques proposed below, is not essential and could be removed whenever the density ϱ does not factorize, again by introducing an auxiliary density $\hat{\varrho} = \frac{1}{2^N}$.

Observe however that this framework does not immediately include problems where $a(\mathbf{x}, \omega)$ is not bounded away from zero, like the important case where $a(\mathbf{x}, \omega)$ is a lognormal random field, i.e. $a(\mathbf{x}, \omega) = e^{\gamma(\mathbf{x},\omega)}$, with $\gamma(\mathbf{x}, \omega)$ being a Gaussian random field.

Finally, we denote by $L^2_\varrho(\Gamma)$ the space of square integrable functions on Γ with respect to the measure $\frac{1}{2^N}d\mathbf{y}$, and by $V = H^1_0(D)$ the space of square integrable functions in D with square integrable distributional derivatives and with zero trace on the boundary, equipped with the gradient norm $\|v\|_V = \|\nabla v\|_{L^2(D)}$, $\forall v \in V$. Its dual space will be denoted by V'. Moreover, since V and $L^2_\varrho(\Gamma)$ are Hilbert spaces, we can define the tensor space $V \otimes L^2_\varrho(\Gamma)$ as the completion of formal sums $v(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{k'} v_{D,k}(\mathbf{x})v_{\Gamma,k}(\mathbf{y})$, $\{v_{D,k}\} \subset V$, $\{v_{\Gamma,k}\} \subset L^2_\varrho(\Gamma)$ with respect to the inner product

$$(v, \widehat{v})_{V\otimes L^2_\varrho(\Gamma)} = \sum_{k,\ell} (v_{D,k}, \widehat{v}_{D,\ell})_V \, (v_{\Gamma,k}, \widehat{v}_{\Gamma,\ell})_{L^2_\varrho(\Gamma)}.$$

We are now in the position to write a weak formulation of (1),

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

$$\int_\Gamma \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, \varrho(\mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x}, \mathbf{y}) \, \varrho(\mathbf{y}) \, d\mathbf{x} \, d\mathbf{y}. \tag{2}$$

Thanks again to Assumption A1 and the Lax–Milgram lemma, there exists a unique solution to problem (2) for any $f \in V'$, with $\|u\|_{V\otimes L^2_\varrho(\Gamma)} \leq \frac{\|f\|_{V'}}{a_{\min}}$. We remark that u can be understood either as a function in the tensor space $H^1_0(D) \otimes L^2_\varrho(\Gamma)$ or as a $H^1_0(D)$-valued square-integrable function of $\mathbf{y} \in \Gamma$, i.e. $u \in L^2_\varrho(\Gamma; H^1_0(D))$; we will use either notation depending on the situation.

2.2. Regularity of u with respect to the random parameters

Concerning the regularity of the solution u with respect to the input $\mathbf{y}$, it is well-known that, under reasonable assumptions on the regularity of the coefficient a, u is analytic in every $\mathbf{y} \in \Gamma$. We refer e.g. to [5,6] for a proof in the case of linear dependence of the diffusion coefficient a on the parameters y_i, and to [3] for the more general case in which $a(\mathbf{x}, \mathbf{y})$ is infinitely many times differentiable with respect to $\mathbf{y}$ and $\exists\, r_1, \ldots, r_N \in \mathbb{R}_+$ s.t.

$$\left\| \frac{1}{a} \cdot \frac{\partial^{i_1+\cdots+i_N} a}{\partial y_1^{i_1} \cdots \partial y_N^{i_N}} \right\|_{L^\infty(D)} \leq \prod_{n=1}^{N} r_n^{i_n} \quad \forall \mathbf{y} \in \Gamma, \ \forall\, i_1, \ldots, i_N \in \mathbb{N}. \tag{3}$$

In this work, we will restrict our focus to the case in which u obeys the following assumption:

Assumption A3 (*"Polydisc Analyticity"*)**.** The complex continuation of u, denoted by $u^* : \mathbb{C}^N \to H^1_0(D)$ is a $H^1_0(D)$-valued holomorphic function in the polydisc

$$E_{S_1,\ldots,S_N} = \prod_{n=1}^{N} E_{n,S_n}, \quad E_{n,S_n} = \{z_n \in \mathbb{C} : |z_n| \leq S_n\}$$

for each $1 < S_n < S_n^*$, with $\sup_{\mathbf{z}\in E_{S_1,\ldots,S_N}} \|u^*(\mathbf{z})\|_{H^1_0(D)} \leq B_u$, and $B_u = B_u(S_1, S_2, \ldots, S_N) \to \infty$ as $S_n \to S_n^*$, $n = 1, \ldots, N$.

Remark 3. We will see in Section 5 that this class of functions includes e.g. the solution of the inclusions tests already investigated in [15], as well as other elliptic problems that depend on few coefficients that can be varied independently from one to another in given intervals. An example is given by elasticity problems with uncertain Young modulus and Poisson ratio. On the other hand, this is not the correct framework for diffusion coefficients that have the form $a(\mathbf{x}, \omega) = \sum_{n=1}^{N} b_n(\mathbf{x}) y_n(\omega)$ with functions b_n with overlapping supports, which will be typically the case for a truncated Karhunen–Loève expansion of a correlated random field.

Problem (2) can be discretized in space by introducing e.g. a finite element discretization with piecewise continuous polynomials over a triangulation $\mathcal{T}_h$ of the physical domain D, $V_h(D) \subset H_0^1(D)$. Such semi-discrete solution will thus belong to $V_h(D) \otimes L^2_\varrho(\Gamma)$, and will feature the same regularity properties of the continuous solution u with respect to the random parameters.

3. Galerkin polynomial approximation in the stochastic dimension

In this section we temporarily drop Assumptions A2 and A3, and briefly review the Galerkin approximation method in the more general setting where u depends on N random parameters supposed to be independent and identically distributed. Of course, since the Galerkin method builds an approximation using global polynomials, it will be effective only if u has some regularity with respect to y_i. Thus, we introduce a polynomial subspace of $L^2_\varrho(\Gamma)$, which we denote by $\mathcal{P}_{\mathrm{w}}(\Gamma)$, and look for a fully discrete solution $u^G_{h,\mathrm{w}} \in V_h(D) \otimes \mathcal{P}_{\mathrm{w}}(\Gamma)$ solving

Problem 4 (*Fully Discrete Weak Formulation*)**.** Find $u^G_{h,\mathrm{w}} \in V_h(D) \otimes \mathcal{P}_{\mathrm{w}}(\Gamma)$ such that $\forall\, v \in V_h(D) \otimes \mathcal{P}_{\mathrm{w}}(\Gamma)$

$$\int_\Gamma \int_D a_N(\mathbf{x}, \mathbf{y}) \nabla u^G_{h,\mathrm{w}}(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y})\, \varrho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x}, \mathbf{y})\, \varrho(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}, \tag{4}$$

with the understanding that the polynomial space $\mathcal{P}_{\mathrm{w}}(\Gamma)$ should be designed to have good approximation properties while having a number of degrees of freedom as low as possible. This is the well known *Stochastic Galerkin formulation* (see e.g. [1,7–10]). In this respect a Tensor Product polynomial space that contains all the N-variate polynomials with maximum degree in each variable lower than a given $\mathrm{w} \in \mathbb{N}$ is not a good choice. Indeed, its dimension grows exponentially fast with the number of random variables N, i.e. $\dim \mathcal{P}_{\mathrm{w}}(\Gamma) = (1 + \mathrm{w})^N$. A valid alternative choice that has been widely used in literature (see e.g. [7,20,21]) is given by the Total Degree polynomial space, that includes those polynomials whose total degree is lower than or equal to w: such space contains indeed only $\binom{N+\mathrm{w}}{N}$ polynomials, which is much lower than $(1 + \mathrm{w})^N$, and still has good approximation properties. A number of possible polynomial spaces has been listed and analyzed e.g. in [15]. One could also introduce anisotropy in the approximation, with the aim to enrich the polynomial space only in those directions of the stochastic space which contribute the most to the total variability of the solution.

To solve problem (4) in practice, it is convenient to endow $\mathcal{P}_{\mathrm{w}}(\Gamma)$ with a $\varrho(\mathbf{y})d\mathbf{y}$-orthonormal basis: to this end we take advantage of the tensor structure of $L^2_\varrho(\Gamma)$ and build the elements of such basis as products of $\varrho_n(y_n)dy_n$-orthonormal polynomials on Γ_n, which we denote as $\{\Psi_{q_n}\}_{q_n \in \mathbb{N}}$:

$$\mathbf{\Psi}_{\mathbf{q}}(\mathbf{y}) = \prod_{n=1}^{N} \Psi q_n(y_n) \quad \mathbf{q} = (q_1,\, q_2,\, \ldots,\, q_n),\ \mathbf{q} \in \mathbb{N}^N. \tag{5}$$

Families of $\varrho_n(y_n)dy_n$-orthonormal polynomials exist for many probability distribution: we recall Legendre polynomials for uniform measures and Hermite polynomials for Gaussian measures (see [21] for the general Askey scheme), for which explicit formulas and computing algorithms are available, see e.g. [22]. As a word of caution, we can note that the work [23] showed that there exist probability measures, such as the lognormal one, which admit a family of orthonormal polynomials that however does not form a basis for $L^2_\varrho(\Gamma)$, i.e. there exist functions in $L^2_\varrho(\Gamma)$ that cannot be approximated with arbitrary precision by linear combinations of such orthonormal polynomials.

To construct general polynomial spaces we introduce a sequence of increasing index sets $\Lambda(\mathrm{w})$, $\mathrm{w} \in \mathbb{N}$, such that

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

each with cardinality M (depending on w), and consider the corresponding polynomial spaces

$$\mathcal{P}_{\mathrm{w}}(\Gamma) = \mathbb{P}_{\Lambda(\mathrm{w})}(\Gamma) = \operatorname{span}\left\{\mathbf{\Psi}_{\mathbf{q}}(\mathbf{y}),\ \mathbf{q} \in \Lambda(\mathrm{w})\right\}$$

for the approximation of $u^G_{h,\mathrm{w}}$ with the Stochastic Galerkin method. In other words, the Stochastic Galerkin method will compute the coefficients $u^G_{\mathbf{q}} \in V_h(D)$ of the expansion

$$u^G_{h,\Lambda(\mathrm{w})}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{q} \in \Lambda(\mathrm{w})} u^G_{\mathbf{q}}(\mathbf{x}) \mathbf{\Psi}_{\mathbf{q}}(\mathbf{y}). \tag{6}$$

Such expansion is usually known as generalized Polynomial Chaos Expansion (gPCE). Having the gPCE expansion of $u^G_{h,\mathrm{w}}$ (6) allows us to compute easily the mean and variance of $u^G_{h,\mathrm{w}}$ as

$$\mathbb{E}\left[u^G_{h,\Lambda(\mathrm{w})}(\mathbf{x},\cdot)\right] = u^G_{\mathbf{0}}(\mathbf{x}), \qquad \mathbb{V}\mathrm{ar}\left[u^G_{h,\Lambda(\mathrm{w})}(\mathbf{x},\cdot)\right] = \sum_{\mathbf{q}\in\Lambda(\mathrm{w})\setminus\Lambda(0)} u^G_{\mathbf{q}}(\mathbf{x})^2.$$

Finally, using (6) in the weak formulation (4) and choosing as test function $v_h(\mathbf{x})\boldsymbol{\Psi}_{\boldsymbol{\kappa}}(\mathbf{y})$, v_h being a finite element basis function, we obtain a set of M linear systems for the modes $u^G_{\mathbf{q}}(\mathbf{x})$, that will be usually coupled due to the presence in (4) of non-zero terms like $\int_{\Gamma_n} a(\mathbf{x},\mathbf{y})\Psi_{q_n}(y_{q_n})\Psi_{\kappa_n}(y_{\kappa_n})\varrho_n(y_n)dy_n$; see e.g. [15,24] and references therein for more details on the discrete problem.

4. Quasi-optimal stochastic Galerkin method for analytic functions in polydiscs

We now go back to the specific case of u satisfying Assumptions A1–A3, and we consider the basis for $\mathbb{P}_{\Lambda(\mathrm{w})}(\Gamma)$ given by multivariate Legendre polynomials. In what follows, we do not consider the approximation in the physical space, and only consider the Galerkin solution $u^G_{\Lambda(\mathrm{w})}$ in the space $V\otimes\mathbb{P}_{\Lambda(\mathrm{w})}(\Gamma)$. We also introduce the truncated Legendre expansion $u_{\Lambda(\mathrm{w})}$ of the exact solution u on $V\otimes\mathbb{P}_{\Lambda(\mathrm{w})}(\Gamma)$,

$$u_{\Lambda(\mathrm{w})}(\mathbf{x},\mathbf{y}) = \sum_{\mathbf{q}\in\Lambda(\mathrm{w})} u_{\mathbf{q}}(\mathbf{x})\boldsymbol{\Psi}_{\mathbf{q}}(\mathbf{y}), \quad u_{\mathbf{q}} = \int_{\Gamma} u(\mathbf{x},\mathbf{y})\boldsymbol{\Psi}_{\mathbf{q}}(\mathbf{y})\varrho(\mathbf{y})d\mathbf{y}. \tag{7}$$

We first recall the following optimality result for the Stochastic Galerkin approximation, whose proof can be found e.g. in [25].

Theorem 5. *Under Assumption A1, we have that the Stochastic Galerkin solution $u^G_{\Lambda(\mathrm{w})}$ corresponding to $\mathbb{P}_{\Lambda(\mathrm{w})}(\Gamma)$ satisfies*

$$\begin{aligned}\left\|u-u^G_{\Lambda(\mathrm{w})}\right\|_{V\otimes L^2_\varrho(\Gamma)} &\le C_{\mathrm{opt}} \inf_{v\in V\otimes\mathbb{P}_{\Lambda(\mathrm{w})}(\Gamma)} \|u-v\|_{V\otimes L^2_\varrho(\Gamma)} \\ &= C_{\mathrm{opt}} \left\|u-u_{\Lambda(\mathrm{w})}\right\|_{V\otimes L^2_\varrho(\Gamma)},\end{aligned}$$

where C_{opt} is a constant depending on $a_{\min}$, $a_{\max}$.

From Theorem 5 we see that the optimal M-dimensional polynomial space for the Stochastic Galerkin method is the one spanned by the Legendre polynomials corresponding to the M largest coefficients in the truncated Legendre expansion (7). This choice indeed minimizes the energy of the projection error

$$\left\|u-u_{\Lambda(\mathrm{w})}\right\|^2_{V\otimes L^2_\varrho(\Gamma)} = \left\|u-\sum_{\mathbf{q}\in\Lambda(\mathrm{w})} u_{\mathbf{q}}\boldsymbol{\Psi}_{\mathbf{q}}\right\|^2_{V\otimes L^2_\varrho(\Gamma)} = \sum_{\mathbf{q}\notin\Lambda(\mathrm{w})} \|u_{\mathbf{q}}\|^2_V,$$

over all the possible choices of $\Lambda(\mathrm{w})$ with fixed cardinality M.

A possible strategy to assess the convergence rate of the resulting approximation of u is to order the Legendre coefficients $\|u_{\mathbf{q}}\|^2_V$ in decreasing order according to a suitable a-priori estimate and study the summability properties of the sequence thus obtained. This idea has been investigated e.g. in [5,6] for the case when the diffusion coefficient can be written as $a(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{\infty} y_i b_i(\mathbf{x})$, with y_i uniform random variables over $[-1,1]$ and $\{\|b_i\|_\infty\}_{i\in\mathbb{N}} \in \ell^p$ for some $p < 1$. It is then possible to prove an algebraic convergence of the L^2_ϱ error with rate $1/p - 1/2$. The proof is however not constructive, i.e. no algorithm is presented to build a sequence of polynomial approximations with such convergence rate. Uniform convergence results are given in [5], as well as in [16,17].

In this work we will restrict our focus to the case in which the solution u obeys Assumption A3. In this case we are able to give explicit formulas for the construction of a sequence of polynomial approximations that is "quasi-optimal" (i.e. optimal with respect to a sharp upper bound of the Legendre coefficients), and to prove a subexponential rate of convergence for such sequence of approximations.

4.1. Construction of the quasi-optimal polynomial space

We start by proving a result on the decay of the coefficients of the Legendre expansion for u satisfying Assumption A3. To this end, we first need the following simple lemma, whose proof is straightforward.

Definition 6. Let $\mathcal{E}_{\delta_1,\dots,\delta_N}$ be the family of Bernstein polyellipses $\mathcal{E}_{\delta_1,\dots,\delta_N} = \prod_{n=1}^N \mathcal{E}_{n,\delta_n}$ with

$$\mathcal{E}_{n,\delta_n} = \left\{ z_n\in\mathbb{C} : \Re\mathfrak{e}\,(z) = \frac{\delta_n+\delta_n^{-1}}{2}\cos\phi,\ \Im\mathfrak{m}\,(z) = \frac{\delta_n-\delta_n^{-1}}{2}\sin\phi,\ \phi\in[0,2\pi)\right\}, \quad \delta_n > 1.$$

Lemma 7. *Let $\delta_n(S_n) = S_n + \sqrt{S_n^2-1}$, with S_n as in Assumption A3. The polyellipse $\mathcal{E}_{\delta_1(S_1),\dots,\delta_N(S_N)}$ is the largest polyellipse of the family $\mathcal{E}_{\delta_1,\dots,\delta_N}$ included in the polydisc $E_{S_1,\dots,S_N}$ in Assumption A3.*

Next, we also need to introduce the monodimensional $L^\infty(\Gamma)$-normalized Legendre polynomials $\Psi_j^\infty(t), j = 0, 1, \ldots$ for which the following properties hold:

- $\Psi_j^\infty(1) = 1$;
- $\Psi_j(t) = \sqrt{2j+1}\Psi_j^\infty(t)$ with $\Psi_j(t)$ as in (5).

We are now in the position to prove the following estimate on the Legendre coefficients.

Proposition 8. *If the solution u fulfills Assumptions A1–A3, the coefficients of the Legendre expansion (7) decay as*

$$\left\|u_{\mathbf{q}}\right\|_V \leq C_{\text{Leg}}\, e^{-\sum_{n=1}^N g_n q_n} \prod_{n=1}^N \sqrt{2q_n+1}, \tag{8}$$

with $g_n = \log(\delta_n(S_n))$ and

$$C_{\text{Leg}}(S_1, \ldots, S_N) = B_u(S_1, \ldots, S_N) \prod_{n=1}^N \frac{l(\mathcal{E}_{n,\delta_n(S_n)})}{4(\delta_n(S_n)-1)},$$

for all $S_n < S_n^*$.

Here $l(\mathcal{E}_{n,\delta_n(S_n)})$ denotes the length of the ellipse $\mathcal{E}_{n,\delta_n(S_n)}$ in Lemma 7, $\delta_n(S_n)$ is as in Lemma 7, and $B_u(S_1, \ldots, S_N)$ is as in Assumption A3.

Proof. The proof follows closely the argument in [14, Section 12.4]. Once we have fixed the radii $S_n < S_n^*$ in Assumption A3, from Lemma 7 we have that u is analytic in and on $\mathcal{E}_{\delta_1(S_1),\ldots,\delta_N(S_N)}$, and hence we can exploit Cauchy's formula to rewrite the $\mathbf{q}$-th Legendre coefficient as

$$\begin{aligned} u_{\mathbf{q}} &= \int_\Gamma u(\mathbf{x},\mathbf{y})\boldsymbol{\Psi}_{\mathbf{q}}(\mathbf{y})\varrho(\mathbf{y})d\mathbf{y} \\ &= \int_\Gamma \boldsymbol{\Psi}_{\mathbf{q}}(\mathbf{y})\varrho(\mathbf{y}) \oint_{\mathcal{E}_{\delta_1(S_1),\ldots,\delta_N(S_N)}} \frac{u^*(\mathbf{x},\mathbf{z})}{\prod_n 2\pi i(z_n - y_n)} d\mathbf{z}d\mathbf{y} \\ &= \oint_{\mathcal{E}_{\delta_1(S_1),\ldots,\delta_N(S_N)}} u^*(\mathbf{x},\mathbf{z}) \prod_{n=1}^N \frac{1}{2}\int_{\Gamma_n} \frac{\Psi_{q_n}(y_n)}{2\pi i(z_n - y_n)} dy_n d\mathbf{z}. \end{aligned}$$

Next, let

$$\mathbb{I}_{q_n}(z_n) = \int_{\Gamma_n} \frac{\Psi_{q_n}^\infty(y_n)}{(z_n - y_n)} dy_n.$$

From [14, Lemma 12.4.6] it follows that for all $z_n \in \mathcal{E}_{n,\delta_n(S_n)}$ we have

$$|\,\mathbb{I}_{q_n}(z_n)| \leq \pi \frac{(1/\delta_n(S_n))^{q_n}}{\delta_n(S_n) - 1}.$$

Then we can estimate the $\mathbf{q}$-th Legendre coefficient of u by

$$\begin{aligned} \left\|u_{\mathbf{q}}\right\|_V &\leq \sup_{\mathcal{E}_{\delta_1(S_1),\ldots,\delta_N(S_N)}} \left\|u^*\right\|_V \prod_{n=1}^N \frac{\sqrt{2q_n+1}}{4\pi} \oint_{\mathcal{E}_{n,\delta_n}} |\,\mathbb{I}_{q_n}(z_n)| dz_n \\ &\leq \sup_{\mathcal{E}_{\delta_1(S_1),\ldots,\delta_N(S_N)}} \left\|u^*\right\|_V \prod_{n=1}^N \frac{\sqrt{2q_n+1}}{4\pi} \pi \frac{(1/\delta_n(S_n))^{q_n}}{\delta_n(S_n)-1} \oint_{\mathcal{E}_{n,\delta_n}} dz_n \\ &\leq \sup_{\mathcal{E}_{\delta_1(S_1),\ldots,\delta_N(S_N)}} \left\|u^*\right\|_V \prod_{n=1}^N \frac{\sqrt{2q_n+1}\, l(\mathcal{E}_{n,\delta_n})}{4(\delta_n(S_n)-1)} e^{-q_n \log(\delta_n(S_n))}. \end{aligned}$$

Finally observe that

$$\sup_{\mathcal{E}_{\delta_1(S_1),\ldots,\delta_N(S_N)}} \left\|u^*\right\|_V \leq \sup_{E_{S_1,\ldots,S_N}} \left\|u^*\right\|_V \leq B_u(S_1, \ldots, S_N). \quad \square$$

Observe that the square root factor in (8) is asymptotically negligible compared to the exponentially decreasing term $e^{-\sum_n g_n q_n}$. Motivated by this fact, we introduce the following corollary, that will be crucial in the following of the paper.

Corollary 9 (*Exponential Decay of the Legendre Coefficients*)**.** *The Legendre coefficients of u satisfying Assumptions A1–A3 can be accurately estimated as*

$$\left\|u_{\mathbf{q}}\right\|_V \leq \widehat{C}_{\text{Leg}} \prod_{n=1}^{N} e^{-\widehat{g}_n q_n} \tag{9}$$

for some $\widehat{g}_n < g_n$ and $\widehat{C}_{\text{Leg}} > C_{\text{Leg}}$. For instance, for all $0 < \epsilon < 1$, one could take $\widehat{g}_n = g_n(1-\epsilon)$ and $\widehat{C}_{\text{Leg}} = C_{\text{Leg}} \prod_n \left(e^{\epsilon g_n/2}/\sqrt{\epsilon g_n e}\right)$.

Given the estimate for the decay of the Legendre coefficients of u in Eq. (9), the family of (anisotropic) Total Degree (*TD*) sets $\mathbb{P}_{TD(\mathrm{w},\widehat{\mathbf{g}})}(\Gamma)$, with

$$TD(\mathrm{w},\widehat{\mathbf{g}}) = \left\{\mathbf{q} \in \mathbb{N}^N : \sum_{n=1}^{N} \widehat{g}_n q_n \leq \mathrm{w}\right\},$$

is a sharp estimate of the optimal polynomial space for the Stochastic Galerkin method, provided that estimate (9) is in turn sharp. Indeed, following the procedure proposed in [3], one can define the quasi-optimal index set Λ by selecting all multiindices $\mathbf{q}$ for which the *estimated decay* of the corresponding Legendre coefficient is above a fixed threshold $\epsilon \in \mathbb{R}_+$,

$$\Lambda_\epsilon = \left\{\mathbf{q} \in \mathbb{N}^N : \widehat{C}_{\text{Leg}} \prod_{n=1}^{N} e^{-\widehat{g}_n q_n} \geq \epsilon\right\},$$

or equivalently

$$\Lambda(\mathrm{w}) = \left\{\mathbf{q} \in \mathbb{N}^N : \sum_{n=1}^{N} \widehat{g}_n q_n \leq \mathrm{w},\ \mathrm{w} = \left\lceil -\log\epsilon/\widehat{C}_{\text{Leg}} \right\rceil\right\} = TD(\mathrm{w},\widehat{\mathbf{g}}).$$

We now derive convergence estimates for $u_{TD(\mathrm{w},\widehat{\mathbf{g}})}$, following different arguments for the isotropic and anisotropic problem case. We anticipate that the numerical tests presented in Section 6 will confirm that (9) is indeed a very sharp estimate of the decay of the Legendre coefficients when $\widehat{g}_n$ are properly tuned, at least for the isotropic case, and the convergence of the resulting *TD* approximation is very close to the convergence of the best M-terms approximation.

4.2. Convergence analysis for the isotropic case

We begin the convergence analysis for the *TD* Galerkin approximation of u by the isotropic setting, following closely the argument in [25]. Therefore, we further assume that Assumption A3 holds with $S_n = S$, for $n = 1, \ldots, N$. As a consequence, the parameters δ_n describing the polyellipses in Lemma 7 are all equal, as well as the coefficients $\widehat{g}_n$ driving the decay of the Legendre coefficients in Proposition 8 and Corollary 9. Thus the optimal polynomial space is indeed the isotropic Total Degree, $TD(\mathrm{w},\mathbf{1}) = \{\mathbf{q} \in \mathbb{N}^N : \sum_{n=1}^{N} q_n \leq \mathrm{w}\}$. For simplicity, we will denote this set simply as $TD(\mathrm{w})$, and the corresponding Galerkin solution as $u^G_{TD(\mathrm{w})}$. Moreover, we will denote the polydiscs in Assumption A3 as E_S, the constant in Assumption A3 as $B_u(S)$ and the polyellipses in Lemma 7 and Proposition 8 as $\mathcal{E}_{\delta(S)}$.

We shall need the following lemma (see [26] for a proof).

Lemma 10. *Suppose that u satisfies Assumption A3 with $S_n = S$ for $n = 1, \ldots, N$, and let $\mathcal{M}_{u,\mathrm{w}}$ be the Maclaurin polynomial of u on the complex domain,*

$$\mathcal{M}_{u,\mathrm{w}}(\mathbf{z}) = \sum_{\mathbf{q} \in TD(\mathrm{w})} \alpha_{\mathbf{q}} \prod_{n=1}^{N} z_n^{q_n},$$

with $\alpha_{\mathbf{q}} \in V$,

$$\alpha_{\mathbf{q}}(\mathbf{x}) = \frac{1}{\prod_{n=1}^{N} q_n!} \frac{\partial^{q_1+\cdots+q_n}}{\partial y_1^{q_1} \cdots \partial y_N^{q_n}} u(\mathbf{x},\mathbf{y})_{|\mathbf{y}=\mathbf{0}}.$$

Then, for any $0 < R < S$, we have the estimate

$$\sup_{\mathbf{z} \in E_R} \left\|u^*(\mathbf{z}) - \mathcal{M}_{u,\mathrm{w}}(\mathbf{z})\right\|_V \leq \frac{B_u(S)}{S/R - 1} e^{-h\mathrm{w}},$$

with $B_u(S)$ as in Assumption A3 and $h = \log\frac{S}{R}$.

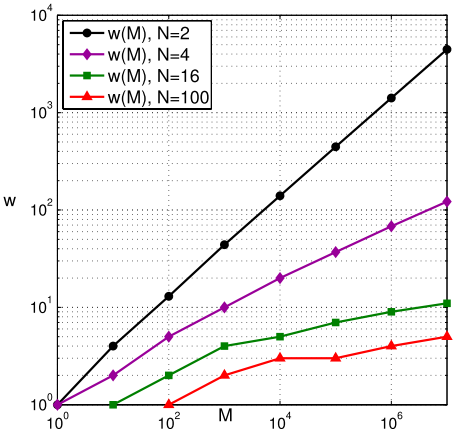

Fig. 1. $\mathrm{w}(M)$ for different values of N.

The convergence rate for the isotropic *TD* Galerkin approximation can then be estimated combining Theorem 5 and Lemma 10.

Theorem 11. *Suppose that u satisfies Assumptions A1–A3 with $S_n = S$ for $n = 1, \dots, N$. Then, the Stochastic Galerkin solution $u_{TD(\mathrm{w})}$ satisfies*

$$\left\|u - u^G_{TD(\mathrm{w})}\right\|_{V\otimes L^2_\varrho(\Gamma)} \le C_{\mathrm{opt}} \frac{B_u(S)}{S-1} e^{-h\mathrm{w}},$$

with $B_u(S)$ as in Lemma 10, $h = \log S$ and C_{opt} as in Theorem 5.

Proof. We use Lemma 10 with $R = 1$ (note that the intersection of E_1 with the real axis is Γ). Then we have

$$\begin{aligned}\left\|u - u^G_{TD(\mathrm{w})}\right\|_{V\otimes L^2_\varrho(\Gamma)} &\le C_{\mathrm{opt}} \inf_{v\in V\otimes\mathbb{P}_{TD(\mathrm{w})}(\Gamma)} \|u - v\|_{V\otimes L^2_\varrho(\Gamma)} \\ &\le C_{\mathrm{opt}} \left\|u - \mathcal{M}_{\mathrm{w},u}\right\|_{V\otimes L^2_\varrho(\Gamma)} \\ &\le C_{\mathrm{opt}} \left\|u - \mathcal{M}_{\mathrm{w},u}\right\|_{L^\infty(\Gamma;V)} \le C_{\mathrm{opt}} \frac{B_u(S)}{S-1} e^{-h\mathrm{w}}. \quad \square\end{aligned}$$

Theorem 11 states an exponential convergence of the error with respect to the total degree of the polynomial approximation. In practice however one is more concerned with the convergence of $u_{TD(\mathrm{w})}$ with respect to the number of degrees of freedom, i.e. the dimension M of the space $TD(\mathrm{w})$. Hence, we are led to the problem of finding an estimate for the function $\mathrm{w} = \mathrm{w}(M)$.

Note that the inverse of such function, $M = M(\mathrm{w})$, is known analytically, $M = \binom{N+\mathrm{w}}{N}$. The function $\mathrm{w}(M)$ could thus be easily computed numerically: it is of course increasing in M and decreasing in N, i.e. the level w needed to have M terms in the set is lower for higher N, see Fig. 1. In general, we can state the following proposition.

Proposition 12. *Under the same hypotheses of Theorem 11, for every $M > 0$ there holds*

$$\left\|u - u^G_{TD(\mathrm{w})}\right\|_{V\otimes L^2_\varrho(\Gamma)} \le C_{\mathrm{opt}} \frac{B_u(S)}{S-1} M^{-h/(1+\log N)}, \tag{10}$$

with $B_u(S)$ as in Lemma 10, $h = \log S$ and C_{opt} as in Theorem 5. Furthermore, in the asymptotic limit $\mathrm{w} \ge N$, that holds for instance if $M > 4^N$, there holds

$$\left\|u - u^G_{TD(\mathrm{w})}\right\|_{V\otimes L^2_\varrho(\Gamma)} \le C_{\mathrm{opt}} \frac{B_u(S)}{S-1} e^{-\frac{hN}{2e}\sqrt[N]{M}}. \tag{11}$$

Proof. Eq. (10) can be proved (see also [25, Eq. (25)]) by observing that

$$\begin{aligned}M = \prod_{i=1}^{N}\left(1 + \frac{\mathrm{w}}{i}\right) &= \exp\left(\sum_{i=1}^{N} \log\left(1 + \frac{\mathrm{w}}{i}\right)\right) \le \exp\left(\sum_{i=1}^{N} \frac{\mathrm{w}}{i}\right) \\ &= \exp\left(\mathrm{w}\sum_{i=1}^{N} \frac{1}{i}\right) \le e^{\mathrm{w}(\log(N)+1)}.\end{aligned}$$

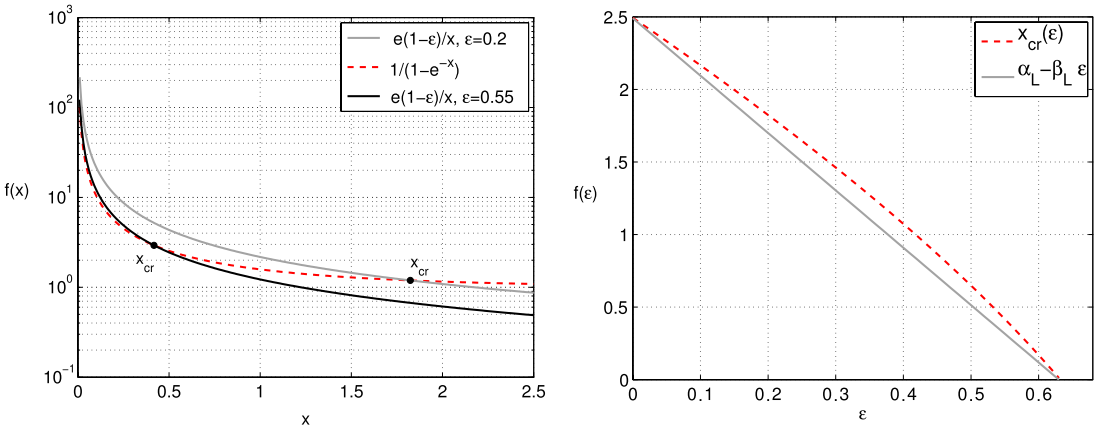

Fig. 2. Left: graphical representation of inequality (12) for $\epsilon = 0.2$ and $\epsilon = 0.55$. Right: value of $x_{\rm cr}$ and bound in Eq. (13).

Therefore $\log M \le \mathrm{w}\,(\log(N)+1)$, hence $\mathrm{w} \ge \frac{\log M}{1+\log N}$ and $e^{-\mathrm{w}h} \le M^{-h/(1+\log N)}$. In the asymptotic limit $\mathrm{w} \ge N$ we have instead

$$M = \prod_{i=1}^{N}\left(1+\frac{\mathrm{w}}{i}\right) \le \frac{2^N \mathrm{w}^N}{N!} \Rightarrow \mathrm{w} \ge \left(N! 2^{-N} M\right)^{1/N} \ge \frac{N}{2e} M^{1/N}.$$

Finally, using the well-known Stirling approximation of $N!$ we have that $\binom{2N}{N} \le 4^N$ for all $N > 0$ so that $M > 4^N$ implies $\mathrm{w} \ge N$. □

4.3. Convergence analysis for the anisotropic case

In this section we remove the isotropic assumption, and we derive a convergence estimate for the Galerkin solution $u^G_{TD(\mathrm{w},\widehat{\mathbf{g}})}$ with an argument substantially different from the previous section. We start with two technical lemmas that we will need in the following.

Lemma 13. *For $0 < \epsilon < \frac{e-1}{e} = \epsilon_{\max} \approx 0.63$, there holds*

$$\frac{1}{1-e^{-x}} \le \frac{(1-\epsilon)e}{x}, \quad 0 < x \le x_{\rm cr}(\epsilon). \tag{12}$$

Moreover, the function $x_{\rm cr}(\epsilon)$ is concave and can be bounded as

$$\alpha_L - \beta_L \epsilon \le x_{\rm cr}(\epsilon) \tag{13}$$

with $\alpha_L \approx 2.49$, $\beta_L = (\alpha_L/\epsilon_{\max})$.

Proof. For $x > 0$ and $\epsilon < 1$, (12) is actually equivalent to

$$e^{-x} \le 1 - \frac{1}{(1-\epsilon)e}x$$

that can hold for $0 < x < x_{\rm cr}(\epsilon)$ only if $-\frac{1}{(1-\epsilon)e} > -1$, hence $\epsilon < \frac{e-1}{e}$. The function $x_{\rm cr}(\epsilon)$ can be easily shown to be concave, and its value at $\epsilon = 0$ can be computed numerically as $\alpha_L = x_{\rm cr}(\epsilon) \approx 2.49$, hence (13). □

Lemma 14. *Given any $C_{\log,M} \in (0, 1/e]$, there holds*

$$M \le e^{C_{\log,M} N \sqrt[N]{M}}, \tag{14}$$

for a sufficiently large M, $M > M_{\log}$. In particular, for $C_{\log,M} = 1/e$ the bound holds for any $M > 0$.

Proof. From the trivial observation that given any $C_{\log,M}$ there holds $\log t \le C_{\log,M}\, t$ for sufficiently large t, we have immediately

$$\frac{1}{N}\log(M) = \log(\sqrt[N]{M}) \le C_{\log,M}\sqrt[N]{M} \Rightarrow \log(M) = C_{\log,M} N \sqrt[N]{M},$$

hence the thesis of the lemma. In particular, $\log t$ and $C_{\log,M}\, t$ are tangent in $t = e$, with $C_{\log,M} = 1/e$. □

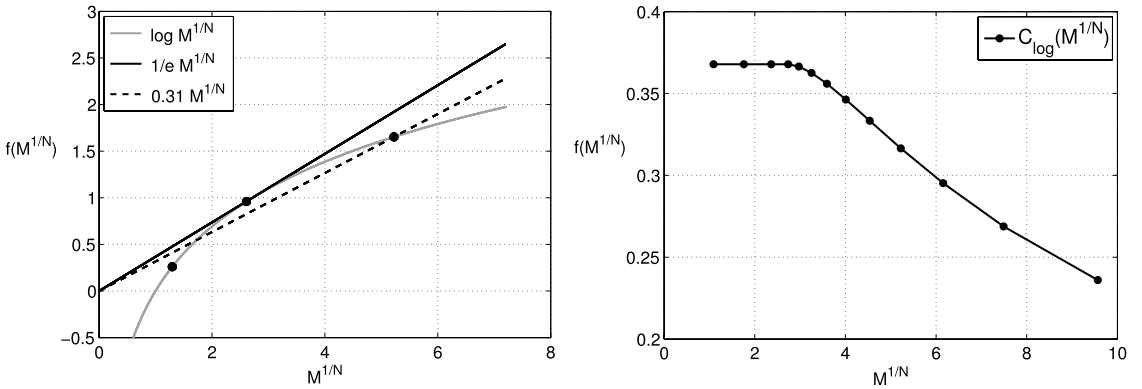

Fig. 3. Graphical representation of Lemma 14. Left: graphical visualization of the bound. The dots represent $M^{1/N} = 0.1, 1/e, 0.5$ respectively. Note that the first point is less than $1/e$, therefore the same bound as $M^{1/N} = 1/e$ applies. Right: visualization of $M^{1/N}$ vs. $C_{\log,M}$ for $1 \leq M^{1/N} \leq 10$, numerically assessed.

Fig. 2-left shows the effectiveness of (12), while Fig. 2-right shows the function $x_{\mathrm{cr}}(\epsilon)$, as well as the bound in Eq. (13). Similarly, Fig. 3-left shows some instances of estimate (14), while Fig. 3-right shows the value of $C_{\log,M}$ corresponding to a range of values of $\sqrt[N]{M}$.

Next, we consider again the expression for the L^2_ϱ projection error

$$\|u - u_{\Lambda(\mathrm{w})}\|^2_{V\otimes L^2_\varrho(\Gamma)} = \left\| u - \sum_{\mathbf{q}\in\Lambda(\mathrm{w})} u_{\mathbf{q}}\boldsymbol{\Psi}_{\mathbf{q}} \right\|^2_{V\otimes L^2_\varrho(\Gamma)} = \sum_{\mathbf{q}\notin\Lambda(\mathrm{w})} \|u_{\mathbf{q}}\|^2_V$$

where $\Lambda(\mathrm{w})$ is now the set of multiindices corresponding to the best M-terms approximation. Having estimated such optimal set with the total degree set $TD(\mathrm{w}, \widehat{\mathbf{g}})$ and the decay of the Legendre coefficients as exponential in each variable, according to Corollary 9, we have that

$$\begin{aligned}\left\| u - \sum_{\mathbf{q}\in\Lambda(\mathrm{w})} u_{\mathbf{q}}\boldsymbol{\Psi}_{\mathbf{q}} \right\|^2_{V\otimes L^2_\varrho(\Gamma)} &\leq \left\| u - \sum_{\mathbf{q}\in TD(\mathrm{w},\widehat{\mathbf{g}})} u_{\mathbf{q}}\boldsymbol{\Psi}_{\mathbf{q}} \right\|^2_{V\otimes L^2_\varrho(\Gamma)} \\ &= \sum_{\mathbf{q}\notin TD(\mathrm{w},\widehat{\mathbf{g}})} \|u_{\mathbf{q}}\|^2_V \leq \widehat{C}^2_{\mathrm{Leg}} \sum_{\mathbf{q}\in\mathbb{N}^N, \mathbf{q}\cdot\widehat{\mathbf{g}}>\mathrm{w}} e^{-2\mathbf{q}\cdot\widehat{\mathbf{g}}},\end{aligned}$$

and we will concentrate on bounding the last term of this inequality, $\sum_{\mathbf{q}\cdot\widehat{\mathbf{g}}>\mathrm{w}} e^{-2\mathbf{q}\cdot\widehat{\mathbf{g}}}$. To this end, we will need the so-called Stechkin Lemma, see e.g. [27].

Lemma 15 (*Stechkin*). *Let $0 \leq p \leq q$, and let $\{a_j\}_{j\in\mathbb{N}}$ be a positive decreasing sequence. Then*

$$\left(\sum_{j>M} (a_j)^q\right)^{1/q} \leq M^{-\frac{1}{p}+\frac{1}{q}} \left(\sum_{j\in\mathbb{N}} (a_j)^p\right)^{1/p}.$$

We are now ready to state the main result of this section.

Theorem 16. *Suppose the Legendre coefficients of u can be bounded as in Corollary 9. Let g_m be the geometric mean of the rates of the decay of the Legendre coefficients, $g_m = \sqrt[N]{\prod_{n=1}^N \widehat{g}_n}$. Consider the level w anisotropic TD approximation of u with rates $\widehat{\mathbf{g}}$, and denote by M its cardinality. Finally, let*

$$\mathbb{S}(\tau) = \left(\sum_{\mathbf{q}\in\mathbb{N}^N} e^{-2\tau\mathbf{q}\cdot\widehat{\mathbf{g}}}\right)^{1/\tau} = \left(\prod_{n=1}^N \frac{1}{1-e^{-2\tau\widehat{g}_n}}\right)^{1/\tau} < \infty$$

for every $\tau > 0$. Then under Assumptions A1–A3 there holds

$$\|u - u^G_{TD(\mathrm{w},\widehat{\mathbf{g}})}\|^2_{V\otimes L^2_\varrho(\Gamma)} \leq C^2_{\mathrm{opt}}\widehat{C}^2_{\mathrm{Leg}} \exp\left(N\sqrt[N]{M}\left(C_{\log,M} - \frac{2g_m\delta}{e}\right)\right), \tag{15}$$

for $0 < \delta < \epsilon_{\max}$, $\epsilon_{\max}$ *as in Lemma 13,* $C_{\log,M}$ *as in Lemma 14 and*

$$M > \left(\frac{\widehat{g}_N e}{g_m(\alpha_L - \delta\beta_L)}\right)^N. \tag{16}$$

Proof. Using the estimate on the Legendre coefficients in Corollary 9 and Lemma 15 with $q = 1$, $p = \tau$, we have for the L^2_ϱ projection

$$\frac{1}{\widehat{C}^2_{\text{Leg}}}\|u - u_{TD(\mathrm{w},\widehat{\mathbf{g}})}\|^2_{V\otimes L^2_\varrho(\Gamma)} \le \sum_{\mathbf{q}\cdot\widehat{\mathbf{g}}>\mathrm{w}} e^{-2\mathbf{q}\cdot\widehat{\mathbf{g}}} \le M^{1-\frac{1}{\tau}}\mathbb{S}(\tau). \tag{17}$$

Now, since (17) holds for every $\tau > 0$ we would like to compute τ^* minimizing $\frac{\mathbb{S}(\tau)}{\sqrt[\tau]{M}}$,

$$\tau^* = \arg\min_{\tau\in\mathbb{R}_+} \frac{\mathbb{S}(\tau)}{\sqrt[\tau]{M}} = \arg\min_{\tau\in\mathbb{R}_+} \left(\frac{1}{M\prod_{n=1}^N (1 - e^{-2\tau\widehat{g}_n})}\right)^{1/\tau}.$$

We do not solve exactly this problem and just discuss the approximated value $\tau^* = e/(2g_m\sqrt[N]{M})$. This choice is motivated in the case $\tau\widehat{g}_n$ small $\forall n = 1,\dots,N$, so that $1 - e^{-2\tau\widehat{g}_n} \approx 2\widehat{g}_n\tau$, as τ^* is the exact optimum solution of the approximated problem

$$\tau^* = \arg\min_{\tau\in\mathbb{R}} \left(\frac{1}{M\tau^N 2^N\prod_{n=1}^N \widehat{g}_n}\right)^{1/\tau}.$$

Plugging $\tau^* = e/(2g_m\sqrt[N]{M})$ in (17) we obtain

$$\sum_{\mathbf{q}\cdot\widehat{\mathbf{g}}>\mathrm{w}} e^{-2\mathbf{q}\cdot\widehat{\mathbf{g}}} \le M\left(\frac{1}{M\prod_{n=1}^N (1 - e^{-\widehat{g}_n e/(g_m\sqrt[N]{M})})}\right)^{2g_m\sqrt[N]{M}/e}. \tag{18}$$

Next we apply Lemma 13 to bound $1/\left(1 - e^{-\widehat{g}_n e/(g_m\sqrt[N]{M})}\right)$, obtaining

$$\frac{1}{1 - e^{-\widehat{g}_n e/(g_m\sqrt[N]{M})}} \le \frac{(1-\epsilon_{M,n})g_m\sqrt[N]{M}}{\widehat{g}_n}, \quad \text{for } \frac{\widehat{g}_n e}{g_m\sqrt[N]{M}} \le x_{\mathrm{cr}}(\epsilon_{M,n}), \tag{19}$$

so that Eq. (18) simplifies to

$$\sum_{\mathbf{q}\cdot\widehat{\mathbf{g}}>\mathrm{w}} e^{-2\mathbf{q}\cdot\widehat{\mathbf{g}}} \le M\left(\prod_{n=1}^N (1-\epsilon_{M,n})\right)^{2g_m\sqrt[N]{M}/e}. \tag{20}$$

By using the lower bound in (13), we see that condition (19)-right holds if we choose $\epsilon_{M,n}$ as

$$\widehat{g}_n e/(g_m\sqrt[N]{M}) = \alpha_L - \beta_L\epsilon_{M,n} \Rightarrow \epsilon_{M,n} = \left(\alpha_L - \frac{\widehat{g}_n e}{g_m\sqrt[N]{M}}\right)\frac{1}{\beta_L}.$$

Moreover, we see from (20) that to ensure convergence of the estimate we need $\epsilon_{M,n} > 0$, which enforces a constraint on M. Namely, taken any $0 < \delta < \epsilon_{\max}$ we require $\epsilon_{M,n} > \delta$ which implies

$$\delta < \left(\alpha_L - \frac{\widehat{g}_n e}{g_m\sqrt[N]{M}}\right)\frac{1}{\beta_L} \Rightarrow M > \left(\frac{\widehat{g}_n e}{g_m(\alpha_L - \delta\beta_L)}\right)^N.$$

See Remark 17 for more details on the choice of δ. Furthermore, note that the rates are supposed to be ordered increasingly, so that this condition has to be checked for $n = N$ only, hence (16). With this choice of $\epsilon_{M,n}$, Eq. (20) further simplifies to

$$\begin{aligned}\sum_{\mathbf{q}\cdot\widehat{\mathbf{g}}>\mathrm{w}} e^{-2\mathbf{q}\cdot\widehat{\mathbf{g}}} &\le M\left(\prod_{n=1}^{N}\left(1-\epsilon_{M,n}\right)\right)^{2g_m\sqrt[N]{M}/e}\\ &= M\exp\left(2g_m\sqrt[N]{M}/e\sum_{i=1}^{N}\log\left(1-\epsilon_{M,n}\right)\right)\\ &\le M\exp\left(-2g_m\sqrt[N]{M}/e\sum_{i=1}^{N}\epsilon_{M,n}\right)\\ &\le M\exp\left(-\frac{2g_mN\sqrt[N]{M}\delta}{e}\right).\end{aligned} \tag{21}$$

Finally, we apply Lemma 14, to obtain

$$\sum_{\mathbf{q}\cdot\widehat{\mathbf{g}}>\mathrm{w}} e^{-2\mathbf{q}\cdot\widehat{\mathbf{g}}} \le \exp\left(N\sqrt[N]{M}\left(C_{\log,M}-\frac{2g_m\delta}{e}\right)\right)$$

and the result follows from Theorem 5. □

Remark 17 (*The Role of δ*)**.** Here we neglect the influence of $C_{\log,M}$ in estimate (15) and further investigate the link between M and δ.

On one hand, choosing a small δ will reduce the minimum cardinality M for the estimate to hold, cf. Eq. (16); in the limit $\delta \to 0$, we have $M \ge \left(\frac{\widehat{g}_n e}{g_m\alpha_L}\right)^N$. In the isotropic case, $\widehat{g}_n = g_m$, estimate (15) is of the same form of estimate (11) in Proposition 12, however under the much milder condition $M \ge (e/\alpha_L)^N \approx 1.09^N$; in a problem with $N = 10$ random variables this would correspond to $M > 3$. On the other hand, $\delta = 0$ in (15) would imply no convergence rate. Conversely, the highest convergence would be obtained setting $\delta = \epsilon_{\max}$ but would be realized only in the limit $M \to \infty$.

Remark 18 (*Recovering the Isotropic Result*)**.** We can also compare this result with the isotropic estimate (11) in Proposition 12. In that case for $M > 4^N$ we had a rate of $hN/(2e)$, which one would obtain with (15) by choosing $\delta = \frac{h}{2g_m}$. Considering e.g. the isotropic problem detailed in next section one could estimate numerically $h \approx 1.5$, $g_m \approx 2$, that would imply

$$M > \left(\frac{e}{\alpha_L - \frac{h}{2g_m}\beta_L}\right)^N \approx (2.7)^N \approx 2800,$$

or assess h, g theoretically in terms of the radii of the Bernstein ellipses and analyticity regions in Proposition 8 and Theorem 11, resulting in $h \approx 0.025$, $g_m \approx 0.22$ and then $M > 1.2^N$.

The main drawback of (15) is that, for anisotropic problems, condition (16) on M is dominated by the largest rate, $\widehat{g}_N$. However, for problems with large variations of $\widehat{g}_n$ the random variables corresponding to high values of $\widehat{g}_n$ will not be added to approximations of u with small cardinality M: therefore, one may think of devising an "adaptive" estimate in which the constraint on M and the convergence rate depend on the active variables only.

Remark 19 (*The Interplay between $C_{\log,M}$ and δ*)**.** We now also investigate through some numerical computations the effect of $C_{\log,M}$ on estimate (15). To this end, let us denote $C_\delta = \frac{g_m\delta}{e}$, so that estimate (15) can be written as

$$\|u - u^G_{TD(\mathrm{w},\widehat{\mathbf{g}})}\|^2_{V\otimes L^2_\varrho(\Gamma)} \le C^2_{\mathrm{opt}}\widehat{C}^2_{\mathrm{Leg}}\exp\left(N\sqrt[N]{M}\left(C_{\log,M}-2C_\delta\right)\right).$$

For simplicity, we will work in an isotropic setting, $g_m = \widehat{g}_n$ for $n = 1, \ldots, N$. We consider a uniform sampling of the admissible values of δ, $0 < \delta < \epsilon_{\max}$: for each of these values we compute the corresponding values of C_δ and of $\sqrt[N]{M}$ according to Eq. (16), i.e. $\sqrt[N]{M} = \frac{\widehat{g}_n e}{g_m(\alpha_L-\delta\beta_L)}$ (note that in the isotropic case $\widehat{g}_n$ and g_m cancel), and finally we compute numerically $C_{\log,M}$ corresponding to such $\sqrt[N]{M}$. By comparing the values of $C_{\log,M}$ and C_δ thus obtained we can see (cf. Table 1) that $C_{\log,M}$ plays a non-negligible role, preventing the estimate to go to zero as $M \to \infty$ for small values of δ. This phenomenon is however alleviated if g_m is higher.

We finally close this section with an alternative estimate, presented here for the isotropic case only. Towards this end, we now present a couple of auxiliary results.

Table 1
Numerical values for $C_{\log,M}$ and C_δ.

δ	$\sqrt[N]{M}$	$C_{\log,M}$	$g_m = 1$		$g_m = 2$	
			$2C_\delta$	Rate	$2C_\delta$	Rate
0	1.09	0.368	0	0.368	0	0.368
0.05	1.19	0.368	0.0368	0.331	0.0736	0.294
0.1	1.3	0.368	0.0736	0.294	0.147	0.221
0.15	1.43	0.368	0.11	0.258	0.221	0.147
0.2	1.6	0.368	0.147	0.221	0.294	0.0736
0.25	1.81	0.368	0.184	0.184	0.368	0
0.3	2.08	0.368	0.221	0.147	0.441	−0.073
0.35	2.45	0.368	0.258	0.11	0.515	−0.147
0.4	2.97	0.366	0.294	0.0722	0.589	−0.222
0.45	3.79	0.352	0.331	0.0205	0.662	−0.311
0.5	5.22	0.316	0.368	−0.0514	0.736	−0.419
0.55	8.4	0.253	0.405	−0.151	0.809	−0.556
0.6	21.5	0.143	0.441	−0.299	0.883	−0.74

Lemma 20. *Let $\widehat{\mathbf{g}} = (\widehat{g}_1, \ldots, \widehat{g}_N)$ be a vector of positive entries. For every $\tau > 0$, define*

$$\mathbb{T}(\tau) = \left(\sum_{\mathbf{q} \in \mathbb{N}^N} e^{-\tau \mathbf{q} \cdot \widehat{\mathbf{g}}} \right)^{1/\tau} = \left(\prod_{n=1}^{N} \frac{1}{1 - e^{-\tau \widehat{g}_n}} \right)^{1/\tau} < \infty, \tag{22}$$

and let $M = \sum_{\mathbf{q} \cdot \widehat{\mathbf{g}} \le \mathrm{w}} 1$. Then

$$e^{-\mathrm{w}} \le \frac{\mathbb{T}(\tau)}{\sqrt[\tau]{M}}, \quad \forall \tau > 0. \tag{23}$$

Proof. We have immediately that $M e^{-\tau \mathrm{w}} \le \sum_{\mathbf{q} \cdot \widehat{\mathbf{g}} \le \mathrm{w}} e^{-\tau \mathbf{q} \cdot \widehat{\mathbf{g}}} \le \mathbb{T}(\tau)^\tau$. □

Lemma 21. *Consider two non negative sequences, $\{a_j\}_{j \in \mathbb{N}}$ monotone decreasing and $\{f_j\}_{j \in \mathbb{N}}$ monotone increasing. Then, for a given $\lambda \in (0, 1)$ and $M > 0$ we have*

$$\sum_{j > M} a_j^2 \le \frac{1}{f_M} \sup_{j > M} \left\{ a_j^{2\lambda} f_j \right\} \sum_{j > M} a_j^{2(1-\lambda)}.$$

Proof. There holds

$$\sum_{j > M} a_j^2 = \sum_{j > M} \left(a_j^2 a_j^{-2\lambda} a_j^{2\lambda} \frac{f_j}{f_j} \right) \le \frac{1}{f_M} \sup_{j > M} \left\{ a_j^{2\lambda} f_j \right\} \sum_{j > M} a_j^{2(1-\lambda)}. \quad \square$$

Theorem 22 (*Alternative Isotropic Estimate*)**.** *Suppose the Legendre coefficients of u can be bounded as in Corollary 9, with $g_n = g$, for $n = 1, \ldots, N$. Consider the level w isotropic TD approximation of u, and denote by M its cardinality. Let $\epsilon_{\max} \approx 0.63$ as in Lemma 13 and suppose that M is sufficiently large, namely that $M > 1.09^N$. Then the estimates*

$$\begin{aligned} \| u - u^G_{TD(\mathrm{w})} \|^2_{V \otimes L^2_\varrho(\Gamma)} &\le C^2_{\mathrm{opt}} \widehat{C}^2_{\mathrm{Leg}} (1 - \exp(-g))^{-N} \exp\left(-\frac{gN}{e} \log((1 - \epsilon(M))^{-1}) \sqrt[N]{M} \right) \\ &\le C^2_{\mathrm{opt}} C(g)^N M^{-g/e \, \log((1-\epsilon(M))^{-1}) \, (1 + 1/2 \, \log(M)/N)} \end{aligned} \tag{24}$$

hold, with $C(g) = \widehat{C}^{2/N}_{\mathrm{Leg}} \, \frac{\exp\left(-g/e \, \log((1-\epsilon(M))^{-1})\right)}{1 - \exp(-g)}$, and

$$\epsilon(M) = \epsilon_{\max} \left(1 - \frac{1.09}{\sqrt[N]{M}} \right). \tag{25}$$

Proof. Let $\mathbb{T}(\tau)$ be as in (22). We use Lemma 21 choosing the sequence of Legendre coefficients ordered in decreasing order as $\{a_j\}_{j \in \mathbb{N}}$, and setting $\lambda = 1/2$ and $f_j = \frac{\sqrt[\tau]{j}}{\mathbb{T}(\tau)}$. Observe that, thanks to Lemma 20, we have $\sup_{j > M} \left\{ a_j f_j \right\} \le 1$ for any $\tau > 0$.

We can thus estimate for the L^2_ϱ projection

$$\begin{aligned}\|u-u_{TD(\mathrm{w})}\|^2_{V\otimes L^2_\varrho(\Gamma)} &= \widehat{C}^2_{\mathrm{Leg}}\sum_{j>M}a_j^2 \le \widehat{C}^2_{\mathrm{Leg}}\min_{\tau>0}\frac{\mathbb{T}(\tau)}{\sqrt[\tau]{M}}\sum_{j>M}a_j\\ &\le \widehat{C}^2_{\mathrm{Leg}}\min_{\tau>0}\frac{\mathbb{T}(\tau)}{\sqrt[\tau]{M}}\sum_{j>0}a_j\\ &\le \widehat{C}^2_{\mathrm{Leg}}\min_{\tau>0}\frac{\mathbb{T}(\tau)}{\sqrt[\tau]{M}}(1-\exp(-g))^{-N}.\end{aligned}$$

Consider as before, for a given value of $\tau>0$, the approximate minimization of $\frac{\mathbb{T}(\tau)}{\sqrt[\tau]{M}}=\left(\frac{1}{M(1-e^{-\tau g})^N}\right)^{1/\tau}$. Taking $\tau=\frac{e}{g\sqrt[N]{M}}$ yields

$$\begin{aligned}\sum_{j>M}a_j^2 &\le \widehat{C}^2_{\mathrm{Leg}}(1-e^{-g})^{-N}\left(\frac{1}{M(1-e^{-e/\sqrt[N]{M}})^N}\right)^{\frac{g\sqrt[N]{M}}{e}}\\ &\le \widehat{C}^2_{\mathrm{Leg}}(1-e^{-g})^{-N}\exp\left(-\frac{gN}{e}\log((1-\epsilon)^{-1})\sqrt[N]{M}\right)\end{aligned}\tag{26}$$

which holds as long as (cf. Lemma 13) $M\ge\left(\frac{e}{\alpha_L-\epsilon\beta_L}\right)^N=\left(\frac{e}{\alpha_L(1-\epsilon\beta_L/\alpha_L)}\right)^N=\left(\frac{e}{\alpha_L(1-\epsilon/\epsilon_{\max})}\right)^N\approx\left(\frac{1.09}{1-\epsilon/0.63}\right)^N>1.09^N$. Observe now that the choice (25) is optimal for the bound (26), and the result follows from Theorem 5. Finally, the last inequality in (24) follows from (26) recalling the inequality $M^{1/N}\ge 1+\frac{\log(M)}{N}+\frac{1}{2}\left(\frac{\log(M)}{N}\right)^2$. □

5. The inclusions problem

We now consider a generic "inclusions problem" in which the diffusion coefficient in (1) is given by

$$a(\mathbf{x},\mathbf{y})=a_0+\sum_{n=1}^{N}\gamma_n\chi_n(\mathbf{x})y_n,\tag{27}$$

where $\chi_n(\mathbf{x})$ are the indicator functions of the disjoint subdomains $D_n\subset D=[0,1]^2$, $D_n\cap D_m=\varnothing$ for $n\neq m$, and y_n are independent random variables uniformly distributed in $[y_{\min},y_{\max}]$ with $y_{\min}>-a_0$, so that Assumptions A1 and A2 are satisfied, as well as condition (3) ensuring the analyticity of u in every $\mathbf{y}\in\Gamma$. Finally, γ_n are real coefficients, $0<\gamma_n\le 1$, whose values determine the possible anisotropy of the problem.

We will first prove that we can apply Corollary 9, and therefore that the *TD* sets are quasi-optimal sets for such problems. Then, we will apply Theorems 11 and 16 and show that the numerical results obtained for such problems are in agreement with the predicted convergence rates.

We shall begin by reparametrizing the diffusion coefficient in terms of new random variables distributed over $[-1,1]$, so that we can apply the discussion of the previous section. For the sake of notation, we will still denote the new variables as y_i, i.e. $y_i\sim\mathcal{U}(-1,1)$. The new diffusion coefficient will be therefore

$$a(\mathbf{x},\mathbf{y})=a_0+\sum_{n=1}^{N}\gamma_n\chi_n(\mathbf{x})\left(\frac{y_n+1}{2}(y_{\max}-y_{\min})+y_{\min}\right).\tag{28}$$

We can now prove the following lemma on the complex analyticity region of u, that we denote by Σ.

Lemma 23. *The complex continuation u^* of the solution u corresponding to a diffusion coefficient (28) is analytic in the region*

$$\Sigma=\prod_{n=1}^{N}\Sigma_n,\quad \Sigma_n=\{z_n\in\mathbb{C}:\mathfrak{Re}(z_n)>T_n\},$$

with $-1>T_n>T_n^*=\frac{2a_0+\gamma_n(y_{\max}+y_{\min})}{\gamma_n(y_{\min}-y_{\max})}$. Moreover, $\sup_{\mathbf{z}\in\Sigma}\|u^*\|_{H^1_0(D)}\le B_u(T)$, with

$$B_u(T)=\frac{\|f\|_{V'}}{a_0+\sum_{n=1}^{N}\gamma_n\left(\frac{1-|T_n|}{2}(y_{\min}-y_{\max})+y_{\min}\right)}.$$

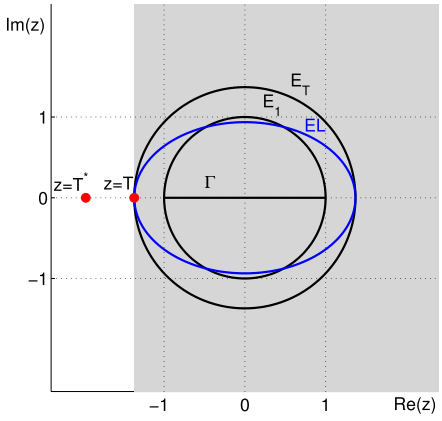

Fig. 4. Regions of the complex plane along the n-th direction for the inclusions problem. For simplicity we drop here the subscript n in the plot. The gray area denotes the analyticity region Σ_n considered. $z_n = T_n^*$ is the singularity up to which it is possible to extend u^* along y_n. *EL* is the ellipse used to estimate the decay of the Legendre coefficients (Proposition 8/Corollary 9), while E_1 and E_T are the circles used to prove the convergence of *TD* estimates in the case of an isotropic setting $\gamma_1, \gamma_2, \ldots, \gamma_N = \gamma$ (Theorem 11).

Proof. As already pointed out, since u satisfies condition (3) then it is analytic in each direction y_n. In particular, having fixed the values of all the random variables but the n-th, let us write $a_n^*(\mathbf{x}, z_n) = a(\mathbf{x}, y_1, y_2, \ldots, y_{n-1}, z_n, y_{n+1}, \ldots, y_N)$ and $u_n^*(\mathbf{x}, z_n) = u(\mathbf{x}, y_1, y_2, \ldots, y_{n-1}, z_n, y_{n+1}, \ldots, y_N)$. Such u_n can be extended in $\Sigma_n = \{z_n \in \mathbb{C} : \Re\mathfrak{e}(z_n) > T_n\}$ for every T_n with $-1 > T_n > T_n^*$, where T_n^* is computed as the value such that

$$\exists\, \mathbf{x} \in D : a_n(\mathbf{x}, T_n^*) = a(\mathbf{x}, y_1, y_2, \ldots, y_{n-1}, T_n^*, y_{n+1}, \ldots, y_N) = 0.$$

This amounts to impose

$$a_0 + \gamma_n \left(\frac{T_n^* + 1}{2} (y_{\max} - y_{\min}) + y_{\min} \right) = 0,$$

whose solution is $T_n^* = (2a_0 + \gamma_n(y_{\max} + y_{\min}))/\big(\gamma_n(y_{\min} - y_{\max})\big)$. Indeed, since the subdomains D_n do not overlap, $a_n(\mathbf{x}, T_n^*) = 0$ in D_n only, i.e. T_n^* does not depend on the value of any of the other random variable y_i. Thus, the analyticity region of u is the Cartesian product of the analyticity regions Σ_n, and the bound for $B_u(T)$ follows immediately. □

5.1. Convergence results

Theorems 11 and 16 apply immediately: in particular, see Fig. 4 for Theorem 11. We summarize the results for the inclusions problem in the following proposition.

Proposition 24. 1. *The Legendre coefficients of the solution of the inclusions problem decay as*

$$\left\| u_{\mathbf{q}} \right\|_V \leq C(\epsilon) e^{-(1-\epsilon)\mathbf{q}\cdot\mathbf{g}} = e^{-\mathbf{q}\cdot\widehat{\mathbf{g}}}, \tag{29}$$

with

$$g_n = \log(|T_n| + \sqrt{T_n^2 - 1}), \quad -1 > T_n > T_n^*, \tag{30}$$

T_n^* as in Lemma 23 and ϵ an in Corollary 9.

2. *The polynomial space $\mathbb{P}_{TD(\mathrm{w},\widehat{\mathbf{g}})}(\Gamma)$ is the quasi-optimal space for the Stochastic Galerkin method when solving the inclusion problem.*
3. *The convergence rate of such quasi-optimal approximation is stated in Theorem 16.*
4. *Moreover, in the isotropic setting where $\gamma_1, \gamma_2, \ldots, \gamma_N = \gamma$, there holds $T_1^* = T_2^* = \cdots = T_N^* = T^*$, $g_1 = g_2 = \cdots = g_N = g$ and we also have an exponential decay of the error with respect to* w *with rate $h = \log|T|$, as stated in Theorem 11.*

In the forthcoming section we will verify the quality of this analysis, both in an isotropic and an anisotropic setting. However, instead of (15) we will actually consider a simplified ansatz, i.e.

$$\|u - u^G_{TD(\mathrm{w},\widehat{\mathbf{g}})}\|^2_{V\otimes L^2_\varrho(\Gamma)} \leq C \exp\left(-\frac{2g_m}{e} N \sqrt[N]{M} \right) \tag{31}$$

and verify that it provides a sharp bound of the error for all $M > 0$.

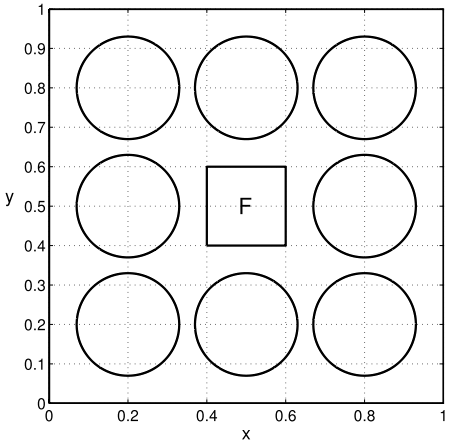

Fig. 5. Physical domain for the isotropic inclusions problem, The inclusions are labeled anti-clock-wise, starting from the bottom-left corner.

6. Numerical results

6.1. Isotropic problem

We now consider the inclusions problem analyzed in [15]. In the first setting considered the subdomains in Eq. (27) are $N = 8$ disjoint circular subdomains as in Fig. 5, $\gamma_n = 1$ for every $n = 1, \ldots, 8$. The random variable y_n are uniformly distributed in $[-0.99, -0.2]$. In addition, we choose $a_0 = 1$ and $f = 100\chi_F$, χ_F being the indicator function of the square located in the middle of the domain, cf. Fig. 5. The aim of this section is to reanalyze the numerical results obtained in [15] in view of the Theorems just proved. In that work, we considered several polynomial approximation spaces, and for each of them we computed the corresponding Stochastic Galerkin approximation, $u^G_{\Lambda(\mathrm{w})}$. Then, we introduced the bounded linear functional $\Theta : H^1_0(D) \to \mathbb{R}$,

$$\Theta(u) = \int_F u(\mathbf{x}) d\mathbf{x}$$

and we monitored the convergence of $\Theta(u^G_{\Lambda(\mathrm{w})})$ with respect to the L^2_ϱ norm error for the Stochastic Galerkin approximation,

$$\varepsilon = \sqrt{\mathbb{E}\left[(\Theta(u^G_{\Lambda(\mathrm{w})}) - \Theta(u_{\mathrm{ref}}))^2\right]}. \tag{32}$$

Note that for this problem we do not have an exact solution, therefore the error is computed with respect to a reference solution. To this end, we have considered the Stochastic Galerkin approximation computed for the *TD* polynomial space at level $\mathrm{w} = 9$, which includes approximately 24 000 Legendre polynomials. The L^2_ϱ error is calculated via a Monte Carlo approximation, i.e.

$$\varepsilon \simeq \left(\frac{1}{W_{MC}} \sum_{l=1}^{W_{MC}} \left[\Theta(u^G_{\Lambda(\mathrm{w})}(\mathbf{y}_l)) - \Theta(u_{\mathrm{ref}}(\mathbf{y}_l))\right]^2\right)^{1/2}, \tag{33}$$

where $\mathbf{y}_l$, $l = 1, \ldots, W_{MC}$, are randomly chosen points in Γ. To this end, $W_{MC} = 1000$ points have proven to be enough to recover a smooth convergence curve.

Fig. 6 shows the effectiveness of the proposed estimate (29) for the decay of the Legendre coefficients in the gPCE expansion of $\Theta(u)$. Indeed, after having computed the Galerkin solution, we have at disposition the coefficients of the gPCE expansion of u, that we can compare with (29). The rates $\widehat{g}$ have been assessed by fitting the Legendre coefficients computed, but the procedure described in [15,3] could have been employed as well. Their numerical value is roughly around 1.90–1.99, i.e. there is no perfect isotropy: this can be explained by the fact that the inclusions are not equally distant from the control area F. Observe that the theoretical rate predicted is at most $\log(|T^*| + \sqrt{T^{*2} - 1}) \approx 0.22$. Thus the estimate we provide in Corollary 9 captures the right behavior of the decay of the Legendre coefficients (i.e. exponential), but is very conservative. Yet, it can still provide the ansatz for a calibrated estimate, which is what we propose in this work.

Fig. 7-left shows the convergence with respect to the level w of the L^2_ϱ error squared for the *TD* approximation of $\Theta(u)$, and shows an optimal agreement between the numerical results and the exponential decay predicted in Theorem 11. Note however that the rate h observed experimentally is $h \approx 1.5$, which is again much larger than the theoretically predicted rate, which amounts to at most $h = \log|T^*| \approx 0.025$.

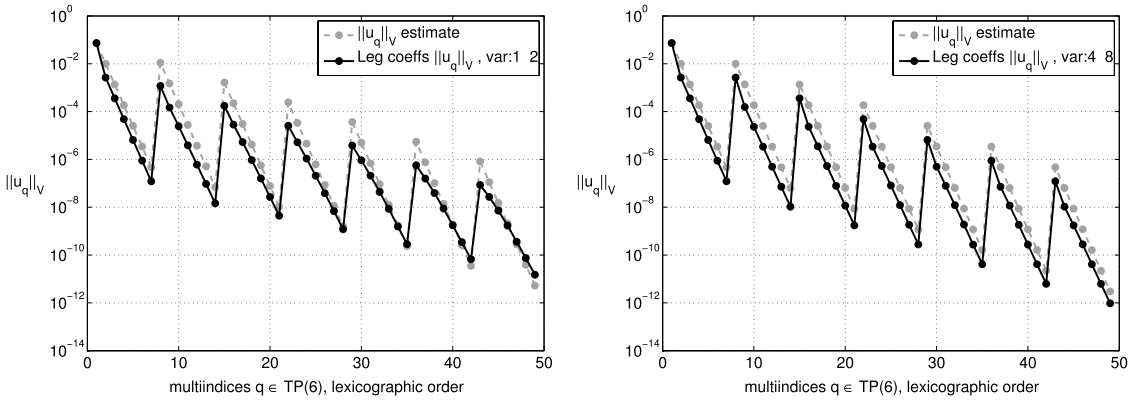

Fig. 6. Comparison between some coefficients of gPCE expansion of $\Theta(u)$, computed with a highly accurate Galerkin approximation (Hyperbolic Cross polynomial space at level $w = 56$, see [15] for the definition) and the corresponding bound (29) suitably tuned. The multiindices corresponding to the coefficients shown in the plots are non-zero only in $y_1 - y_2$ (left) and $y_4 - y_8$ (right) and ordered in lexicographic order.

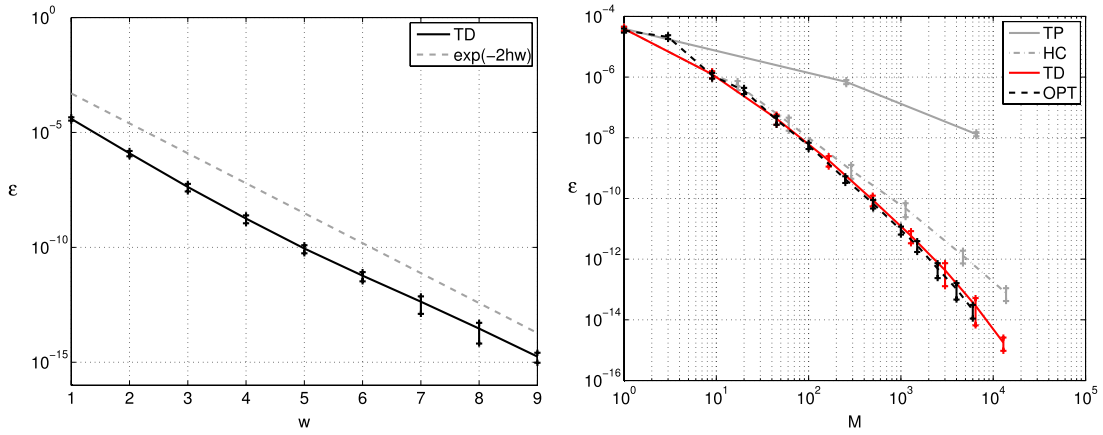

Fig. 7. Left: convergence of the error (33) squared with respect to w for the quasi-optimal TD Galerkin approximation. Right: convergence of the error (33) squared in terms of the dimension of the polynomial space, for the TD approximation, as well as Tensor Product (TP), Hyperbolic Cross (HC) and best M-terms (OPT) approximations.

Fig. 7-right shows instead the convergence with respect to M of the error (33) squared for different polynomial approximations, (namely: Total Degree (TD), Hyperbolic Cross (HC) and Tensor Product (TP) spaces) as well as an estimate of the optimal convergence for the Galerkin method. The latter has been estimated by rearranging in decreasing order the coefficients of the Galerkin solution $TD(9)$ and using again a Monte Carlo estimate for the L^2_ϱ error squared, as in Eq. (33). Since the convergences have been estimated using a Monte Carlo sampling, we also provide in the plot uncertainty bars corresponding to ± 3 standard deviations of the Monte Carlo estimator. As already observed in [15], the TD approximation is the most efficient approximation scheme for the problem of interest, and now can be also understood as the quasi-optimal approximation, as indeed its convergence curve is very close to the best M-terms convergence.

Finally, Fig. 8 shows that the theoretical convergence estimates for the error of the TD approximation appears to be quite sharp, even in its simplified form (31) and apparently without any constraint on M. In particular, observe that the value of g_m used here is 1.9, i.e. it has been computed by fitting the Legendre coefficients (and pretending a perfect isotropy) and not by fitting the error convergence itself (as it was done for Fig. 7 instead). For large values of M however, such simplified estimate seems to be too optimistic. Yet, one should also consider that the convergence curve may be slightly miscalculated, due to the Monte Carlo approximation of the L^2_ϱ error, and to the fact that the Legendre coefficients computed are not exact, but rather approximated by a "overkilling" Galerkin procedure.

6.2. Anisotropic problem

The second test we consider is an anisotropic problem with 4 random variables uniformly distributed in $[-0.99, 0]$, acting on the inclusions illustrated in Fig. 9, located at the corners of the domain. The anisotropy is given by the coefficients γ_n in the expression (27) of the diffusion coefficient, that have been chosen as detailed in Fig. 9.

In contrast with the isotropic setting just analyzed, here the forcing term and the quantity of interest $\Theta(u)$ are now defined over the whole domain rather than on the smaller area F. Finally, the reference solution is now an isotropic TD

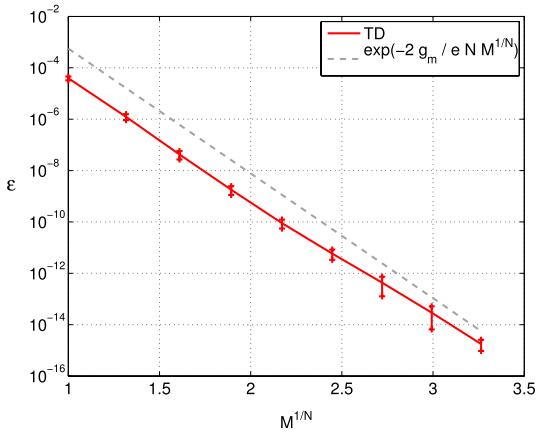

Fig. 8. Convergence of the error (33) squared for the *TD* approximation with respect to $\sqrt[N]{M}$, $N = 8$, compared with the simplified theoretical estimate (31).

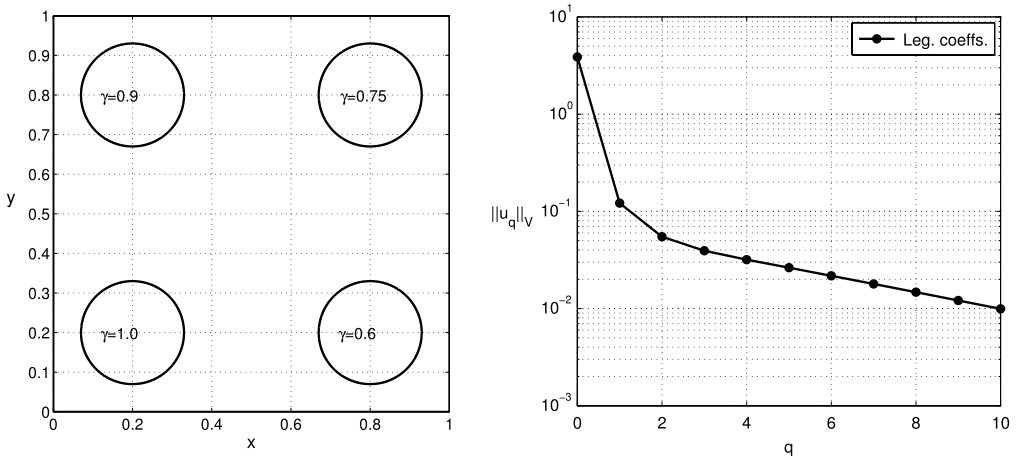

Fig. 9. Left: physical domain for the anisotropic inclusions problem. The numbers inside each inclusion are the corresponding values γ_n. Right: decay of the Legendre coefficients for $\mathbf{q} = [q\,0\,0\,0]$, $0 \le q \le 10$ in semilog scale. A preasymptotic non-exponential regime is clearly present for $q \le 2$.

Stochastic Galerkin approximation at level $\mathrm{w} = 22$, and the L^2_ϱ approximation error is computed with $M = 3000$ Monte Carlo samples.

Compared to the previous case, in this setting the exponential bound on the decay of the Legendre coefficients is not sharp, as a slower preasymptotic regime appears, see Fig. 9-right: in turn this implies that the anisotropic *TD* sets will not be a tight estimate of the best M-terms approximation, see Fig. 10-left. However, using the numerical procedure described in [15,3] it is possible to compute some "effective" exponential rates that yield to anisotropic TD sets with good convergence properties, cf. again Fig. 10-left.

The numerical value of such effective rates is approximately $\widehat{\mathbf{g}} = (0.4,\ 1.37,\ 2.27,\ 3.17)$. Observe that we could also have determined $\widehat{\mathbf{g}}$ by formula (30) in Proposition 24. This would have resulted in $\widehat{\mathbf{g}} \approx (0.20,\ 0.68,\ 1.12,\ 1.51)$, that is roughly half the numerically assessed rates. This is a further confirmation that the theoretical estimates, although not sharp, give a good ansatz to the qualitative features of the problem. Incidentally, note that for the purpose of building a sequence of *TD* sets what really matters is not the absolute value of $\widehat{\mathbf{g}}$, rather the ratio between the rates, the absolute value being important only in the estimate of the convergence rate.

Finally, Fig. 10-right shows that also in this case the simplified estimate (31) on the convergence of the anisotropic *TD* set seems to be quite sharp and to hold without restrictions on the cardinality M of the approximation.

7. Conclusions

In this work we have analyzed the approximability of the solution of linear elliptic PDEs with stochastic coefficients that are analytic in a polydisc in the complex domain. Although somehow restrictive, this hypothesis is satisfied by a number of problems that arise in various engineering fields, as briefly illustrated in Remark 3. This setting has allowed us to use in a very natural way Bernstein ellipses to estimate the decay of the Legendre coefficients, as recalled in Proposition 8, and

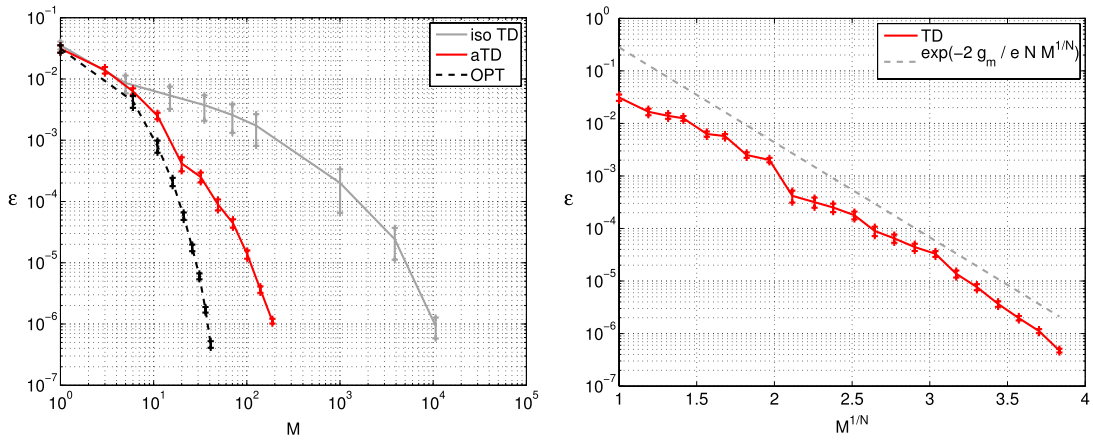

Fig. 10. Left: convergence of the error (33) squared with respect to M for the isotropic and anisotropic *TD* sets, and for the best M terms (OPT) approximations. Right: convergence of the error (33) for the anisotropic *TD* approximation with respect to $\sqrt[N]{M}$, compared with the simplified theoretical estimate (31).

consequently to prove that total degree polynomial spaces represent a quasi-optimal approximation of the best M-terms polynomial approximation. We have then proved with two different arguments the subexponential convergence of the Galerkin approximation of u in such polynomial spaces, see Theorems 11 and 16.

We have verified both the estimate of the decay of the Legendre coefficients and that of the error convergence on two numerical tests, re-examining the results we had obtained in the previous work [15]. The results obtained allow us to claim that the theoretical estimates provided in this work are in essence correct, in the sense that they provide a valid ansatz to be fitted with numerical a-posteriori information, i.e. with a view to a combined a-priori/a-posteriori approach, as already explored in [15,3].

Acknowledgments

The authors would like to recognize the support of the PECOS center at ICES, University of Texas at Austin (Project Number 024550, Center for Predictive Computational Science). Support from the VR project "Effektiva numeriska metoder för stokastiska differentialekvationer med tillämpningar" and King Abdullah University of Science and Technology (KAUST) AEA project "Predictability and Uncertainty Quantification for Models of Porous Media" is also acknowledged. The second and third authors have been supported by the Italian grant FIRB-IDEAS (Project n. RBID08223Z) "Advanced numerical techniques for uncertainty quantification in engineering and life science problems". The fourth author is a member of the KAUST SRI Center for Uncertainty Quantification in Computational Science and Engineering.

References

[1] I. Babuška, R. Tempone, 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, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM Rev. 52 (2010) 317–355.
[3] J. Beck, F. Nobile, L. Tamellini, R. Tempone, On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods, Math. Models Methods Appl. Sci. (M3AS) 22 (2012).
[4] M. Bieri, R. Andreev, C. Schwab, Sparse tensor discretization of elliptic sPDEs, SIAM J. Sci. Comput. 31 (2010) 4281–4304.
[5] A. Cohen, R. Devore, C. Schwab, Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's, Anal. Appl. (Singap.) 9 (2011) 11–47.
[6] A. Cohen, R. DeVore, C. Schwab, Convergence rates of best n-term Galerkin approximations for a class of elliptic sPDEs, Found. Comput. Math. 10 (2010) 615–646.
[7] R.G. Ghanem, P.D. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer-Verlag, New York, 1991.
[8] O.P. Le Maître, O.M. Knio, Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics, in: Scientific Computation, Springer, New York, 2010.
[9] H.G. Matthies, A. Keese, Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations, Comput. Methods Appl. Mech. Engrg. 194 (2005) 1295–1331.
[10] R.A. Todor, C. Schwab, Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients, IMA J. Numer. Anal. 27 (2007) 232–261.
[11] B. Ganapathysubramanian, N. Zabaras, Sparse grid collocation schemes for stochastic natural convection problems, J. Comput. Phys. 225 (2007) 652–685.
[12] F. Nobile, R. Tempone, C. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data, SIAM J. Numer. Anal. 46 (2008) 2309–2345.
[13] D. Xiu, J. Hesthaven, High-order collocation methods for differential equations with random inputs, SIAM J. Sci. Comput. 27 (2005) 1118–1139.
[14] P. Davis, Interpolation and Approximation, Dover Publications Inc., New York, 1975. Republication, with minor corrections, of the 1963 original, with a new preface and bibliography.

[15] J. Bäck, F. Nobile, L. Tamellini, R. Tempone, Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison, in: J. Hesthaven, E. Ronquist (Eds.), Spectral and High Order Methods for Partial Differential Equations, in: Lecture Notes in Computational Science and Engineering, vol. 76, Springer, 2011, pp. 43–62. Selected papers from the ICOSAHOM '09 conference, June 22–26, Trondheim, Norway.
[16] A. Chkifa, A. Cohen, R. DeVore, C. Schwab, Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs, ESAIM Math. Model. Numer. Anal. 47 (2012) 253–280.
[17] C. Gittelson, Uniformely convergent adaptive methods for parametric operator equations, SAM-Report 2011–19, Seminar für Angewandte Mathematik, ETH, Zurich, 2011.
[18] R.B. Nelsen, An Introduction to Copulas, second ed., in: Springer Series in Statistics, Springer, New York, 2006.
[19] I. Babuška, F. Nobile, R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM J. Numer. Anal. 45 (2007) 1005–1034.
[20] G. Stefanou, The stochastic finite element method: past, present and future, Comput. Methods Appl. Mech. Engrg. 198 (2009) 1031–1051.
[21] D. Xiu, G. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2002) 619–644.
[22] W. Gautschi, Orthogonal Polynomials: Computation and Approximation, Oxford University Press, Oxford, 2004.
[23] O.G. Ernst, A. Mugler, H.J. Starkloff, E. Ullmann, On the convergence of generalized polynomial chaos expansions, ESAIM Math. Model. Numer. Anal. 46 (2012) 317–339.
[24] M.F. Pellissetti, R.G. Ghanem, Iterative solution of systems of linear equations arising in the context of stochastic finite elements, Adv. Eng. Softw. 31 (2000) 607–616.
[25] F. Nobile, R. Tempone, Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients, Internat. J. Numer. Methods Engrg. 80 (2009) 979–1006.
[26] T. Bagby, L. Bos, N. Levenberg, Multivariate simultaneous approximation, Constr. Approx. 18 (2002) 569–577.
[27] R.A. DeVore, Nonlinear approximation, in: Acta Numer., vol. 7, Cambridge Univ. Press, Cambridge, 1998, pp. 51–150.