MOX–Report No. 22/2008

Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients

Fabio Nobile, Raul Tempone

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 29 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients

F. Nobile[1], R. Tempone[2,3]

[1] MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, via Bonardi 9, 20133 Milano, Italy
fabio.nobile@polimi.it

[2]Department of Scientific Computing
Florida State University, 400 Dirac Science Library, Tallahassee, FL 32306–4120
rtempone@fsu.edu

[3] Dahlquist Research Fellow, School of Computer Sciences and Communication
KTH, S–100 44 Stockholm, Sweden

Abstract

We consider the problem of numerically approximating statistical moments of the solution of a time dependent linear parabolic partial differential equation (PDE), whose coefficients and/or forcing terms are spatially correlated random fields. The stochastic coefficients of the PDE are approximated by truncated Karhunen-Loève expansions driven by a finite number of uncorrelated random variables. After approximating the stochastic coefficients the original stochastic PDE turns into a new deterministic parametric PDE of the same type, the dimension of the parameter set being equal to the number of random variables introduced.

After proving that the solution of the parametric PDE problem is analytic with respect to the parameters, we consider global polynomial approximations based on tensor product, total degree or sparse polynomial spaces and constructed by either a Stochastic Galerkin or a Stochastic Collocation approach. We derive convergence rates for the different cases and present numerical results that show how these approaches are a valid alternative to the more traditional Monte Carlo Method for this class of problems.

Keywords: PDEs with random data, parabolic equations, multivariate polynomial approximation, Stochastic Galerkin methods, Stochastic Collocation methods, sparse grids, Smolyak approximation, Point Collocation, Monte Carlo Sampling.

AMS subject classification: 65N30, 65N35, 65N12, 65N15, 65C20

1 Introduction

Mathematical models are widely used in many engineering applications to predict the behavior of complex systems, upon which important decisions may be made.

Nowadays, there is an increasing interest in including uncertainty in these models and quantify its effect on the predicted quantities of interest for applications. Such uncertainty may reflect, on the one side, our ignorance or inability to properly characterize all input parameters of the mathematical model; on the other side, it may describe an intrinsic variability of the physical system.

Probability theory offers a natural framework to describe uncertainty, where all uncertain inputs are treated as random variables or more generally random fields. The latter are particularly useful to characterize random spatial variability of some physical properties with a given correlation structure. Examples are the variability of soil permeability in subsurface aquifers, heterogeneity of materials with microstructure, wall roughness in a fluid dynamics study, etc.

Monte Carlo Sampling (MCS) is probably the most natural and widely used technique to forward propagate the input randomness onto the system response or specific quantities of interest. It consists in generating independent realizations drawn from the input distribution and compute sample statistics of the corresponding output values. While being very flexible and easy to implement, MCS features a very slow convergence and does not exploit the possible regularity that the solution might have with respect to the input parameters.

It is known, indeed, that for certain classes of problems, such regularity could be very high. This is the case for the solution of a steady state linear diffusion equation, which actually depends *analytically* on the diffusion coefficient (see e.g. [1, 3, 9]). The high regularity that the solution may have with respect to the input parameters opens up the possibility to use "deterministic" approximations of the *response function* (i.e. the solution of the problem as a function of the input parameters) based on global polynomials, which are expected to yield a very fast convergence and exploit optimally the smoothness of the solution.

When the uncertain input parameters are described by means of second order random fields, the conceptual steps to follow are

1. Represent the random field as a truncated expansion depending on a *finite number of random variables* $Y_1, \dots, Y_N$. This can be done via Fourier-like expansions of the so called Karhunen-Loève expansion (see e.g. [12, Section 3.9]);

2. Having now a finite dimensional noise, the original stochastic problem can be recast into a deterministic parametric one, the random variables acting as parameters;

3. Denoting by $u(Y_1, \dots, Y_N)$ the solution of the problem as a function of the random variables, compute a polynomial approximation $u_p(Y_1, \dots, Y_N)$;

4. Compute statistics of $u_p(Y_1, \ldots, Y_N)$ or some functionals of u_p.

Among these techniques, we mention the so called *Stochastic Galerkin* approach, (see e.g. [11, 21, 33, 2, 28]) which consists in projecting the original equation onto a polynomial subspace which could be constructed as the span of tensor product polynomials or polynomials having total degree smaller that a given integer p. More recently, *Stochastic Collocation* approaches have been proposed (see e.g. [27, 20, 32, 3, 23, 22, 10]) as alternative to Stochastic Galerkin. They consist in computing the solution of the problem in a given set of knots in the parameter space and use these values to construct a global polynomial approximation (sometimes interpolation). The set of knots can be chosen as a tensor or a sparse grid (for instance by following the Smolyak construction [26]).

Clearly, other forms of approximation other than global polynomials are possible and have been explored in recent years, such as wavelet approximations [17, 19] or piecewise polynomials [1, 30]. However, global polynomial approximations are particularly attractive in those cases where the solution features very high, possibly analytic, dependence with respect to the input parameters. We point out that the regularity of the solution is highly problem-dependent and only few results are available so far concerning regularity of the solution and convergence rates for approximating schemes. We mention, for instance, the works [1, 3, 28, 23, 22], which deal with an elliptic equation with a stochastic diffusion coefficient and either Stochastic Galerkin or Collocation approximations, and the work [5], which deals with an elliptic equation defined in a random domain.

In this work, we focus on a linear parabolic PDE with a random diffusivity coefficient described by means of a truncated Karhunen-Loève expansion. We first demonstrate that the solution, interpreted as a Banach-valued function of the input parameters, is analytic as in the steady state case. This result could be proved extending the arguments used in [3]. However, we propose here an alternative proof which is based on analyzing the parabolic equation in the complex plane and checking the Cauchy-Riemann conditions. We also characterize sharply the size and shape of the analyticity region in the complex plane. The proposed technique is quite general and could be of help also in other applications.

We then review Stochastic Galerkin approximations based on either tensor product or total degree polynomial spaces as well as Stochastic Collocation approaches based on either tensor or sparse grids. In all cases, we are able to derive convergence results relying on the regularity result mentioned earlier.

This paper focuses only on approximation techniques with respect to the random parameters. All the analysis is carried out assuming that no time discretization is introduced. In context of a Stochastic Galerkin approximation, the choice of a proper time discretization scheme that results in efficient and stable algorithms is a very important issue and will be addressed in a future work. See e.g. [15] for efficient time marching schemes applied to CFD problems.

We conclude the paper with a numerical example concerning the heat equa-

tion with a random diffusion field having a relatively large correlation length, so that it can be truncated by a relatively small number of random variables (7 in our case). We compare several approaches, namely: Stochastic Galerkin on anisotropic tensor product polynomial spaces; Stochastic Collocation on isotropic sparse grids; Monte Carlo Sampling; as well as the so called *Point Collocation* approach proposed in [14], which is also a global polynomial approximation.

In all cases we observe that the techniques based on global polynomial approximations outperform the Monte Carlo method, at least for a problem with a relatively small number of random variables.

2 Mathematical setting

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

$$\begin{cases} \partial_t u(\omega,t,\mathbf{x}) - \nabla \cdot [a(\omega,\mathbf{x})\nabla u(\omega,t,\mathbf{x})] = f(\omega,t,\mathbf{x}), \text{ in } \Omega \times [0,T] \times D, \\ u(\omega,t,\mathbf{x}) = 0, \text{ on } \Omega \times [0,T] \times \partial D, \\ u(\omega,0,\mathbf{x}) = u_0, \text{ on } \Omega \times D. \end{cases} \tag{1}$$

Here, the symbol ∇ means differentiation with respect to the space variable $\mathbf{x} \in D$. Moreover, $a : \Omega \times D \to \mathbb{R}$ and $f : \Omega \times [0,T] \times D \to \mathbb{R}$ are random functions with continuous and bounded covariance functions. If we denote by $B(D)$ the Borel σ-algebra generated by the open subsets of D, then a and f are assumed measurable with respect to the σ-algebras $(\mathcal{F} \otimes B(D))$ and $(\mathcal{F} \otimes B([0,T] \times D))$, respectively.

Besides, in order to guarantee existence and uniqueness for the solution of (1) we assume that the diffusion coefficient a is bounded and uniformly coercive, i.e.

$$\exists\, a_{\min}, a_{\max} \in (0,+\infty) : \quad P\big(\, \omega \in \Omega :\ a(\omega,\mathbf{x}) \in [a_{\min}, a_{\max}], \quad \forall \mathbf{x} \in \overline{D}\, \big) = 1 \tag{2}$$

and the right hand side in (1) satisfies

$$\int_\Omega \int_{[0,T]\times D} f^2(\omega,t,\mathbf{x}) dx\, dt\, dP < +\infty \tag{3}$$

which implies $\int_{[0,T]\times D} f^2(\omega,t,\mathbf{x}) dx\, dt < +\infty$ almost surely.

Further, to ensure regularity of the solution u with respect to x we assume also that a is globally Lipschitz in $\Omega \times D$.

2.1 Weak formulation

Here we set some basic notation and recall the notion of weak solution. Let Y be an $\mathbb{R}^N$-valued random variable in $(\Omega, \mathcal{F}, P)$ and, for $q \in [1, \infty)$, let $(L_P^q(\Omega))^N$ be the set comprising those random variables Y with $\sum_{i=1}^N \int_\Omega |Y_i(\omega)|^q dP(\omega) < \infty$. If $Y \in L_P^1(\Omega)$ we denote its expected value by

$$\mathbb{E}[Y] = \int_\Omega Y(\omega) dP(\omega) = \int_{\mathbb{R}^N} \mathbf{y}\, d\mu_Y(\mathbf{y}),$$

where μ_Y is the distribution measure for Y, defined for the Borel sets $\tilde{b} \in B(\mathbb{R}^N)$, by $\mu_Y(\tilde{b}) \equiv P(Y^{-1}(\tilde{b}))$. If μ_Y is absolutely continuous with respect to the Lebesgue measure then there exists a density function $\rho : \mathbb{R} \to [0, +\infty)$, such that

$$\mathbb{E}[Y] = \int_{\mathbb{R}^N} \mathbf{y}\, \rho(\mathbf{y})\, d\mathbf{y}.$$

Analogously, whenever $Y \in (L_P^2(\Omega))^N$, the positive semi-definite covariance matrix of Y, $\mathrm{Cov}[Y] \in \mathbb{R}^{N\times N}$, is defined by $\mathrm{Cov}[Y](i,j) = \mathrm{Cov}(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])]$, for $i, j = 1, \ldots, N$. Similarly, for a stochastic function $u = u(\omega, \mathbf{x})$ with $\omega \in \Omega$ and $x \in \overline{D}$, we denote its covariance function by $\mathrm{Cov}[u](\mathbf{x}, \mathbf{x}') = \mathrm{Cov}(u(\cdot, \mathbf{x})\, u(\cdot, \mathbf{x}'))$ for $\mathbf{x}$, $\mathbf{x}' \in \overline{D}$.

Some of our arguments use the notion of the dual space. Let H be a Hilbert space with inner product $(\cdot, \cdot)_H$. The dual space H' of H, contains linear bounded functionals, $\mathcal{L} : H \to \mathbb{R}$, and is endowed with the operator norm $\|\mathcal{L}\|_{H'} = \sup_{v \in H\setminus\{0\}} \frac{\mathcal{L}(v)}{\|v\|_H}$. Besides, the Banach space $C(\Gamma; H)$ comprises all continuous functions $u : \Gamma \to H$ with the norm $\|u\|_{C(\Gamma;H)} \equiv \sup_{y\in\Gamma} \|u(\mathbf{y})\|_H$. Similarly we define (cf. [8, p.285])

$$L_\mu^2(\Gamma; H^k(D)) = \left\{ v : \Gamma \times D \to \mathbb{R} \;\middle|\; v \text{ is strongly meas. and } \int_\Gamma \|v(\mathbf{y}, \cdot)\|_{H^k(D)}^2 d\mu(\mathbf{y}) < +\infty \right\}.$$

We omit the subscript "μ" whenever we refer to the Lebesgue measure.

Since stochastic functions have intrinsically different structure with respect to ω and with respect to x, the analysis of numerical approximations requires tensor spaces. Let H_1, H_2 be Hilbert spaces. The tensor space $H_1 \otimes H_2$ is the completion of formal sums $u = \sum_{i=1}^n v_i\, w_i$, where $\{v_i\}_{i=1}^n \subset H_1$ and $\{w_i\}_{i=1}^n \subset H_2$, with respect to the inner product $(u, \widehat{u})_{H_1\otimes H_2} = \sum_{i,j} (v_i, \widehat{v}_j)_{H_1} (w_i, \widehat{w}_j)_{H_2}$.

We now recall the notion of weak solution for the problem (1): we say that u is a weak solution if it satisfies the initial condition, $u = u_0$ at $t = 0$, and $u \in L^2(0, T; H_0^1(D)) \otimes L_P^2(\Omega)$, $\partial_t u \in L^2(0, T; H^{-1}(D)) \otimes L_P^2(\Omega)$, and a.e. in $[0, T]$

$$\int_D \mathbb{E}[\partial_t u\, v] dx + \int_D \mathbb{E}[a \nabla u \cdot \nabla v] dx = \int_D \mathbb{E}[f v] dx, \quad \forall v \in H_0^1(D) \otimes L_P^2(\Omega). \quad (4)$$

By means of energy estimates, assumptions (2) and (3) imply ([8, Chapter 7]) that there exists a unique solution u in the Hilbert space $H \equiv L^2(0,T;H_0^1(D)) \otimes L_P^2(\Omega)$, endowed with the inner product $(v_1,v_2)_H \equiv \int_{[0,T]\times D} \mathbb{E}[\nabla v_1 \cdot \nabla v_2] dxdt$.

Moreover, the following energy estimate holds

$$\|u(T)\|^2_{L^2(D)\otimes L_P^2(\Omega)} + a_{min}\|u\|^2_{L^2([0,T];H_0^1(D))\otimes L_P^2(\Omega)} \leq \frac{C_D^2}{a_{min}}\|f\|^2_{L^2([0,T]\times D)\otimes L_P^2(\Omega)} + \|u_0\|^2_{L^2(D)} \quad (5)$$

where C_D is the Poincaré constant satisfying: $\|v\|_{L^2(D)} \leq C_D\|\nabla v\|_{L^2(D)}$, for all $v \in H_0^1(D)$.

The main goal in this work is to approximate statistical moments of the solution u or some related quantity of physical interest depending on u. [1]

2.2 Karhunen-Loève expansion and finite dimensional noise

Here we recall the Karhunen-Loève expansion for the approximation of random functions. Consider a random function a with continuous covariance function, $\mathrm{Cov}[a] : \overline{D} \times \overline{D} \to \mathbb{R}$. Let $\{(\lambda_n, b_n)\}_{n=1}^{\infty}$ denote the sequence of eigenpairs associated with the compact self adjoint operator that maps

$$g \in L^2(D) \mapsto \int_D \mathrm{Cov}[a](\mathbf{x},\cdot)g(\mathbf{x})dx \in C^0(\overline{D}).$$

Its non-negative eigenvalues satisfy $\sqrt{\int_{D\times D}(\mathrm{Cov}[a](\mathbf{x}_1,\mathbf{x}_2))^2\, dx_1\, dx_2} \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and $\sum_{n=1}^{+\infty}\lambda_n = \int_D \mathrm{Var}[a](\mathbf{x})dx$. The corresponding eigenfunctions are orthonormal, i.e. $\int_D b_i(\mathbf{x})b_j(\mathbf{x})dx = \delta_{ij}$. The truncated Karhunen-Loève expansion of the random function a, cf. [16], is

$$a_N(\omega,\mathbf{x}) = \mathbb{E}[a](\mathbf{x}) + \sum_{n=1}^{N}\sqrt{\lambda_n}b_n(\mathbf{x})Y_n(\omega) \quad (6)$$

where the real random variables, $\{Y_n\}_{n=1}^{\infty}$, are mutually uncorrelated, have mean zero and unit variance. Whenever $\lambda_n > 0$ these random variables are uniquely determined by $Y_n(\omega) = \frac{1}{\sqrt{\lambda_n}}\int_D (a(\omega,\mathbf{x}) - \mathbb{E}[a](\mathbf{x}))b_n(\mathbf{x})dx$. Then, by Mercer's theorem (cf. [25, p. 245]) we have

$$\sup_{\mathbf{x}\in D}\mathbb{E}[(a-a_N)^2](\mathbf{x}) = \sup_{\mathbf{x}\in D}(\mathrm{Var}[a] - \mathrm{Var}[a_N])(\mathbf{x}) \to 0, \quad \text{as} \quad N \to \infty.$$

If, in addition, the following assumptions are satisfied (see [9]):

[1] Throughout the paper we will assume the initial condition, u_0, and the load f to be deterministic. This assumption may seem restrictive but is not: it can be connected with the more general assumption on the independence of initial conditions and load from the diffusivity coefficient a.

- the images $Y_n(\Omega)$, $n = 1, \ldots,$ are uniformly bounded in $\mathbb{R}$,
- the eigenfunctions b_n are smooth, which is the case when the covariance function is smooth,
- and the eigenpairs have at least the decay $\sqrt{\lambda_n}\|b_n\|_{L^\infty(D)} = \mathcal{O}(\frac{1}{1+n^s})$ for some $s > 1$,

then $\|a - a_N\|_{L^\infty(\Omega\times D)} \to 0$. Notice that for larger values of the decay exponent s we can also obtain the convergence of higher spatial derivatives of a_N in $L^\infty(\Omega \times D)$. The last two conditions can be readily verified once the covariance function of a is known.

Assumption 1 (Finite dimensional noise) *In what follows we assume that the random functions $a(\omega, \mathbf{x})$ and $f(\omega, t, \mathbf{x})$ depend only on an N dimensional random vector Y –this is for instance the case when we use a joint N term Karhunen-Loève expansion to approximate the given coefficients–, i.e. $a(\omega, \mathbf{x}) = a(Y(\omega), \mathbf{x})$ and $f(\omega, t, \mathbf{x}) = f(Y(\omega), t, \mathbf{x})$. Besides, the components of Y, $\{Y_n\}_{n=1}^N$ are uncorrelated real random variables with mean value zero, unit variance, and their images, $\Gamma_n \equiv Y_n(\Omega)$ are bounded intervals in $\mathbb{R}$ for $n = 1, \ldots, N$. Moreover, we assume the vector Y to have a bounded joint probability density function $\rho : \Gamma = \prod_{n=1}^N \Gamma_n \to \mathbb{R}^+$.*

It is usual to have f and a to be independent, because the loads and the material properties are seldom related. In such a situation we have $a(Y(\omega), \mathbf{x}) = a(Y_a(\omega), \mathbf{x})$ and $f(Y(\omega), t, \mathbf{x}) = f(Y_f(\omega), t, \mathbf{x})$, with $Y = [Y_a, Y_f]$ and the vectors Y_a, Y_f independent.

After making Assumption 1, we have by Doob-Dynkin's lemma, cf. [24], that u, the solution corresponding to the stochastic partial differential equation (1) can be described by just a finite number of random variables, i.e. $u(\omega, t, \mathbf{x}) = u(Y_1(\omega), \ldots, Y_N(\omega), t, \mathbf{x})$. Then, in (4), we can replace the probability space $(\Omega, \mathcal{F}, P)$ with $(\Gamma, B(\Gamma), \rho(\mathbf{y})\, d\mathbf{y})$ involving only the image set $\Gamma \subset \mathbb{R}^N$ and the distribution measure for the vector Y, i.e. $d\mu_Y = \rho(\mathbf{y}) d\mathbf{y}$.

This leads to the equivalent formulation

$$\begin{aligned} &\text{Find } u \in L^2(0,T; H_0^1(D)) \otimes L^2_\rho(\Gamma) \text{ with } \partial_t u \in L^2(0,T; H^{-1}(D)) \otimes L^2_\rho(\Gamma), \\ &u|_{t=0} = u_0 \text{ and } \forall v \in H_0^1(D) \otimes L^2_\rho(\Gamma) \text{ and a.e. on } [0,T] \\ &\int_{\Gamma\times D} \partial_t u(t)\, v dx\, \rho d\mathbf{y} + \int_\Gamma \mathcal{B}(u(t), v)\, \rho d\mathbf{y} = \int_{\Gamma\times D} f(t) v dx\, \rho d\mathbf{y}, \end{aligned} \tag{7}$$

with the notation

$$\mathcal{B}(v_1, v_2)(\mathbf{y}) \equiv \int_D a(\mathbf{y}, \mathbf{x}) \nabla v_1(\mathbf{y}, \mathbf{x}) \cdot \nabla v_2(\mathbf{y}, \mathbf{x}) dx, \quad \forall v_1, v_2 \in H_0^1(D) \otimes L^2_\rho(\Gamma).$$

3 Analyticity of the solution with respect to the random inputs

In this section we prove that the solution u is analytic with respect to each input variable y_n, when the diffusion coefficient has the expression (6), namely it is represented as a truncated Karhunen-Loève expansion. Consider the $n-th$ input variable, y_n. Denote by $\hat{\Gamma}_n$ the set of the remainder input variables, that is $\hat{\Gamma}_n = \prod_{1\leq m\leq N, m\neq n}\Gamma_n$ and let $\hat{\mathbf{y}}_n \in \hat{\Gamma}_n$ be an arbitrary point. We will focus on the first direction only, since the proof for the other directions is analogous. Consequently, let $\Gamma_1 = (y_{min}, y_{max})$ and set $\bar{y} = \frac{y_{min}+y_{max}}{2}$ and $|\Gamma_1| = (y_{max} - y_{min})$. We consider the map $\Psi : [-1,1] \to L^2(0,T;H_0^1(D))$ defined by

$$\Psi(s) = u(y_1(s), \hat{\mathbf{y}}_1, \cdot) \in L^2(0,T;H_0^1(D)) \tag{8}$$

with the affine transformation, $y_1 : [-1,1] \to \Gamma_1$, $y_1(s) \equiv \bar{y} + \frac{|\Gamma_1|}{2}\, s$.

Lemma 1 (Complex continuation) *The function $\Psi : [-1,1] \to L^2(0,T;H_0^1(D)) \cap C^0(0,T;L^2(D)$ can be analytically continuated to the circle of the complex plane*

$$\Sigma(r_1) \equiv \{\eta \in \mathbb{C},\ |\eta| \leq 1 + r_1\}, \qquad \textit{with} \quad r_1 = \frac{a_{min}}{|\Gamma_1|\sqrt{\lambda_1}\|b_1\|_{L^\infty(D)}}. \tag{9}$$

Moreover, the complex-valued function $\Psi(\eta)$ satisfies the estimate

$$\begin{aligned} \|\Psi(\eta,T,\cdot)\|^2_{L^2(D)} + \frac{a_{min}}{2}\|\Psi(\eta)\|^2_{L^2(0,T;H_0^1(D))} \\ \leq \frac{2C_D^2}{a_{min}}\|f\|^2_{L^2(0,T;L^2(D))} + \|u_0\|^2_{L^2(D)}, \end{aligned} \tag{10}$$

for all $\eta \in \Sigma(r_1)$ and $\hat{\mathbf{y}}_1 \in \hat{\Gamma}_1$, with C_D being the Poincaré constant for the domain D.

Proof. Consider the natural extension of the real valued variable s to the complex variable $\eta = s + iw$ in (8). Then, the real valued function $\Psi(s)$ has a natural extension to the complex plane as $\Psi(\eta) = u(y_1(\eta), \hat{\mathbf{y}}_1, \cdot)$ and solves the complex problem

$$\begin{cases} \partial_t\Psi(\eta,\cdot) - \nabla\cdot(a(y_1(\eta),\hat{\mathbf{y}}_1,\cdot)\nabla\Psi(\eta,\cdot)) = f(\cdot) \text{ in } [0,T]\times D, \\ \Psi(\eta,\cdot) = 0 \text{ on } [0,T]\times\partial D, \\ \Psi(\eta,\cdot) = u_0(\cdot) \text{ on } \{t=0\}\times D. \end{cases} \tag{11}$$

If we write $\Psi = \Psi_R + i\Psi_I$ and similarly $a = a_R + ia_I$ and introduce the real valued vector $\mathbf{\Psi} = [\Psi_R, \Psi_I]^T$, then problem (11) is equivalent to the 2×2 system of real equations

$$\partial_t\mathbf{\Psi}(\eta,\cdot) - \nabla\cdot(A\nabla\mathbf{\Psi}(\eta,\cdot)) = \mathbf{f} \tag{12}$$

where we have denoted by $A = \begin{bmatrix} a_R & -a_I \\ a_I & a_R \end{bmatrix}$ and $\mathbf{f} = [f, 0]^T$ since we are assuming the forcing term to be deterministic. The system has initial condition $\mathbf{\Psi}(\eta, 0, \cdot) = [u_0, 0]^T$ and homogeneous Dirichlet boundary conditions.

Problem (12) (or equivalently (11)) admits a unique solution as long as $\min_{x\in D} a_R(\eta)$ is strictly positive. We now show that $\min_{x\in D} a_R(\eta) \geq \frac{a_{min}}{2}$, for all $\eta \in \Sigma(r_1)$ and $\hat{y}_1 \in \hat{\Gamma}_1$. Indeed, we have

$$\begin{aligned} a_R(\eta) = \operatorname{Re} a(y_1(\eta), \hat{\mathbf{y}}_1, \mathbf{x}) &= a(0, \hat{\mathbf{y}}_1, \mathbf{x}) + \sqrt{\lambda_1} b_1(\mathbf{x})(\bar{y} + \frac{|\Gamma_1|}{2} s) \\ &= \{\text{setting } s = t(1 + r_1), \text{ with } t \in [-1, 1]\} \\ &= a(y_1(t), \hat{\mathbf{y}}_1, \mathbf{x}) + \frac{\sqrt{\lambda_1} b_1(\mathbf{x})|\Gamma_1|}{2} t r_1 \\ &\geq a_{min} - \frac{\sqrt{\lambda_1}\|b_1\|_{L^\infty(D)}|\Gamma_1|}{2} r_1 \geq a_{min}/2 \end{aligned} \tag{13}$$

Let $\epsilon > 0$. If we multiply (12) by $\mathbf{\Psi}(\eta, \cdot)$, integrate over D and use (13) we obtain the energy estimate

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\mathbf{\Psi}(\eta)\|^2_{L^2(D)} + \frac{a_{min}}{2}\|\nabla\mathbf{\Psi}(\eta)\|^2_{L^2(D)} &\leq \int_D \mathbf{f}\cdot\mathbf{\Psi}(\eta) \\ &\leq \frac{1}{2\epsilon}\|f\|^2_{L^2(D)} + \frac{\epsilon}{2}\|\mathbf{\Psi}(\eta)\|^2_{L^2(D)} \\ &\leq \frac{1}{2\epsilon}\|f\|^2_{L^2(D)} + \frac{\epsilon C_D^2}{2}\|\nabla\mathbf{\Psi}(\eta)\|^2_{L^2(D)} \end{aligned}$$

with C_D being the Poincaré constant for the domain D. Taking $\epsilon = \frac{a_{\min}}{2C_D^2}$ and integrating in time over $[0, T]$ leads to estimate (10).

Finally, to prove that the complex function $\Psi(\eta)$ is analytic in the strip $\Sigma(r_1)$, we verify the Cauchy-Riemann conditions. By formally differentiating system (12) with respect to $\operatorname{Re}\eta = s$ and $\operatorname{Im}\eta = w$ we obtain

$$\begin{cases} \partial_t\partial_s\Psi_R - \nabla\cdot(a_R\nabla\partial_s\Psi_R - a_I\nabla\partial_s\Psi_I) = \nabla\cdot(\partial_s a_R\nabla\Psi_R - \partial_s a_I\nabla\Psi_I) \\ \partial_t\partial_s\Psi_I - \nabla\cdot(a_I\nabla\partial_s\Psi_R + a_R\nabla\partial_s\Psi_I) = \nabla\cdot(\partial_s a_I\nabla\Psi_R + \partial_s a_R\nabla\Psi_I) \end{cases}$$
$$\begin{cases} \partial_t\partial_w\Psi_R - \nabla\cdot(a_R\nabla\partial_w\Psi_R - a_I\nabla\partial_w\Psi_I) = \nabla\cdot(\partial_w a_R\nabla\Psi_R - \partial_w a_I\nabla\Psi_I) \\ \partial_t\partial_w\Psi_I - \nabla\cdot(a_I\nabla\partial_w\Psi_R + a_R\nabla\partial_w\Psi_I) = \nabla\cdot(\partial_w a_I\nabla\Psi_R + \partial_w a_R\nabla\Psi_I) \end{cases}$$

Hence, the derivatives $\partial_s\mathbf{\Psi}$ and $\partial_w\mathbf{\Psi}$ exist everywhere in $\Sigma(r_1)$. Moreover, it is easy to see that the two functions $\Theta(\eta) = \partial_s\Psi_R(\eta) - \partial_w\Psi_I(\eta)$ and $\Xi(\eta) = \partial_w\Psi_R(\eta) + \partial_s\Psi_I(\eta)$ satisfy the system

$$\begin{cases} \partial_t\Theta - \nabla\cdot(a_R\nabla\Theta - a_I\nabla\Xi) = \nabla\cdot((\partial_s a_R - \partial_w a_I)\nabla\Psi_R - (\partial_w a_R + \partial_s a_I)\nabla\Psi_I) \\ \partial_t\Xi - \nabla\cdot(a_I\nabla\Theta + a_R\nabla\Xi) = \nabla\cdot((\partial_w a_R + \partial_s a_I)\nabla\Psi_R + (\partial_s a_R - \partial_w a_I)\nabla\Psi_I) \end{cases} \tag{14}$$

Since the coefficient $a(\eta) = a(0, \hat{\mathbf{y}}_1, \mathbf{x}) + \sqrt{\lambda_1} b_1(\mathbf{x}) \left(\bar{y} + \frac{|\Gamma_1|}{2} \eta \right)$ is linear in η and therefore satisfies the Cauchy-Riemann conditions, the right hand side in (14) vanishes. Finally, we observe that (14) admits the only solution $\Theta(\eta) = \Xi(\eta) = 0$, for all $\eta \in \Sigma(r_1)$ and this proves the analyticity of $\Psi(\eta)$.

□

Remark 1 *Similarly, it can be proved that $\partial_t \Psi : [-1, 1] \to L^2(0, T; H^{-1}(D))$ is also analytic in $\Sigma(r_1)$. Moreover, if we let $\hat{a}(\eta) = \max_{x \in D} |a(\eta, x)|$ then we have*

$$\begin{aligned} \frac{\|\partial_t \Psi\|^2_{L^2([0,T];H^{-1}(D))}}{2} \leq & \hat{a}^2(\eta) \|\Psi\|^2_{L^2([0,T];H^1_0(D))} + C_D^2 \|f\|^2_{L^2([0,T];L^2(D))} \\ \leq & 2 \left(\left(\frac{2\hat{a}(\eta)}{a_{min}} \right)^2 + 1 \right) C_D^2 \|f\|^2_{L^2([0,T];L^2(D))} + \frac{4\hat{a}^2(\eta)}{a_{min}} \|u_0\|^2_{L^2(D)}. \end{aligned}$$

4 Stochastic Galerkin approximation based on polynomial spaces

Let us consider a Galerkin approximation for (7): choose a suitable finite dimensional approximating space, $V_{p,h} \subset H^1_0(D) \otimes L^2_\rho(\Gamma)$, and for each $0 < t < T$ find $u_{p,h}(t) \in V_{p,h}$ such that

$$\int_{\Gamma \times D} \partial_t u_{p,h}(t)\, v\, dx\, \rho d\mathbf{y} + \int_\Gamma \mathcal{B}(u_{p,h}(t), v)\, \rho d\mathbf{y} = \int_{\Gamma \times D} f(t) v\, dx\, \rho d\mathbf{y}, \quad \forall v \in V_{p,h}, \tag{15}$$

with the initial condition $u_{p,h}(0) = u_{0,h}$, being $u_{0,h}$ a suitable projection of the initial condition onto the discrete space $V_{p,h}$. This approximation is understood as a semidiscretization of (7) because it does not carry out a time discretization and yields a system of ordinary differential equations, just as with the classical method of lines.

Here the construction of the subspace $V_{p,h}$ is based on a tensor product, $V_{p,h} = H_h(D) \otimes \mathcal{P}_p(\Gamma)$, where

- $H_h(D) \subset H^1_0(D)$ is a standard finite element space that contains continuous piecewise polynomials defined on regular triangulations τ_h that have a maximum mesh spacing parameter $h > 0$ and
- $\mathcal{P}_p(\Gamma) \subset L^2(\Gamma)$ is a suitably chosen polynomial subspace which is indexed on a parameter p.

For instance, we have the following classical choices:

Example 1 (Anisotropic tensor polynomials) *Let* $\mathbf{p}$ *be a multi-index,* $\mathbf{p} = (p_1, \ldots, p_N)$*. Here the subspace* $\mathcal{P}_{\mathbf{p}}(\Gamma)$ *is the span of tensor product polynomials with degree at most* $\mathbf{p} = (p_1, \ldots, p_N)$ *in each direction, i.e.*

$$\mathcal{P}_{\mathbf{p}}(\Gamma) = \bigotimes_{n=1}^{N} \mathcal{P}_{p_n}(\Gamma_n), \tag{16}$$

with

$$\mathcal{P}_{p_n}(\Gamma_n) = \Big\{ v \in L^2(\Gamma_n) : \quad v \in \text{span}(y^m,\, m = 0, \ldots, p_n) \Big\}, \quad n = 1, \ldots, N.$$

Clearly, the dimension of this subspace is $\eta(\mathbf{p}) = \prod_{n=1}^{N}(1 + p_n)$*.*

Example 2 (Total degree polynomials) *Let* p *be a natural number and let* $\mathcal{P}_p(\Gamma)$ *be the span of monomials with total degree at most* p*, i.e.*

$$\mathcal{P}_p(\Gamma) = span(\prod_{n=1}^{N} y_n^{i_n} : \sum_{n=1}^{N} i_n \leq p). \tag{17}$$

Observe that the dimension of this subspace is $\eta(p, N) = \frac{(N+p)!}{N!p!}$*.*

It is possible to consider other approximation spaces for $\mathcal{P}_p(\Gamma)$. For instance one may use piecewise polynomial functions [7, 1, 29, 30, 28], wavelets [17, 18, 19], etc.

If $\{\varphi_j\}_{j=1}^{N_h}$ and $\{\psi_k\}_{k=1}^{\eta}$ are basis for $H_h(D)$ and $\mathcal{P}_p(\Gamma)$, respectively, we can express the approximate solution as

$$u_{p,h}(\mathbf{y}, t, \mathbf{x}) = \sum_{k=1}^{\eta} \sum_{j=1}^{N_h} u_{kj}(t)\varphi_j(\mathbf{x})\psi_k(\mathbf{y}). \tag{18}$$

The unknown time dependent functions $u_{kj} : [0, T] \to \mathbb{R}$, $k = 1, \ldots, N_h$, $j = 1, \ldots, \eta$ solve a system of ordinary differential equations. This system can readily be obtained by substituting the ansatz (18) into (15) and testing with the basis functions of $V_{p,h}$. Observe that such system has a dimension $N_h * \eta$. Formulation (15) with the choice of finite dimensional spaces introduced above will be referres to as SGFEM.

4.1 Optimality of Stochastic Galerkin approximations

The goal of this section is to analyze the Stochastic Galerkin approximate solution introduced in (15) and derive its optimality in the proper energy norm. This fact is crucial to prove later the rate of convergence of the method.

For instance, Stochastic Galerkin based on tensor product of polynomials yields an exponential rate of convergence with respect to p, the degree of the

polynomials used for approximation, cf. Theorem 2. The application of the p-version in the y direction is motivated by the fact that u is analytic with respect to $\mathbf{y} \in \Gamma$, as we showed in Section 3.

For the purposes of the analysis, we first introduce an auxiliary semi-discrete solution, u_p, satisfying (15) with $\hat{V}_{p,h} = H_0^1(D) \otimes \mathcal{P}_p(\Gamma)$ instead of $V_{p,h} = H_h(D) \otimes \mathcal{P}_p(\Gamma)$.

In other words, for each $0 < t < T$ ones finds $u_p(t) \in H_0^1(D) \otimes \mathcal{P}_p(\Gamma)$ such that

$$\int_{\Gamma \times D} \partial_t u_p(t)\, v\, dx\, \rho d\mathbf{y} + \int_{\Gamma} \mathcal{B}(u_p(t), v)\, \rho d\mathbf{y} = \int_{\Gamma \times D} f(t) v\, dx\, \rho d\mathbf{y}, \quad \forall v \in V_{p,h}, \tag{19}$$

with the exact initial condition $u_p(0) = u_0$.

With this definition, u_p does not have spatial nor time discretization, and we can just concentrate on the $L^2_\rho(\Gamma)$ approximation. Indeed, we have the error splitting

$$\underbrace{u - u_{p,h}}_{\text{full } p\text{-}h \text{ discretization error}} = \underbrace{u - u_p}_{p\text{-version error}} + \underbrace{u_p - u_{p,h}}_{\text{space FEM discretization error}}.$$

Therefore, let us now estimate the p-version error, $e_p \equiv u - u_p$. To this end, we first prove an optimality result for the semidiscrete solution u_p and then use the analyticity of u to achieve exponential convergence in the p-version error. To make the presentation simpler we present only the case of a deterministic forcing term and a stochastic diffusivity coefficient. Since the solution of the parabolic equation, u depends linearly on the forcing f our results generalize directly to the case of a stochastic forcing term as well, provided that the stochastic forcing is an $L^2_\rho(\Gamma)$ valued analytic function of the inputs.

Theorem 1 (Optimality of the SGFEM approximation) *Let $u(\mathbf{y}, \cdot)$ be the solution to (7) and $u_p(\mathbf{y}, \cdot)$ its semi-discrete approximation defined by (19). Consider a function $w \in L^2(0,T; \mathcal{P}_p(\Gamma) \otimes H_0^1(D))$ with $\partial_t w \in L^2(0,T; \mathcal{P}_p(\Gamma) \otimes H^{-1}(D))$ such that $u(0,\cdot) = u_p(0,\cdot) = w(0,\cdot)$. Then we have the estimate*

$$\begin{aligned}
\frac{1}{4}\mathbb{E}[\|(u_p - u)(T,\cdot)\|^2_{L^2(D)}] + \frac{a_{\min}}{4}\mathbb{E}[\|u_p - u\|^2_{L^2(0,T;H_0^1(D))}] \\
\leq \frac{1}{2}\mathbb{E}[\|(u - w)(T,\cdot)\|^2_{L^2(D)}] + \frac{1}{a_{\min}}\mathbb{E}[\|\partial_t(u - w)\|^2_{L^2(0,T;H^{-1}(D))}] \\
+ \left(\frac{a^2_{\max}}{a_{\min}} + \frac{a_{\min}}{2}\right)\mathbb{E}[\|u - w\|^2_{L^2(0,T;H_0^1(D))}],
\end{aligned} \tag{20}$$

Proof.

Let us first recall that

$$\mathbb{E}[< \partial_t u, v > + \mathcal{B}(u, v)] = \mathbb{E}[\int_D f\, v], \qquad \forall v \in H_0^1(D) \otimes L^2_\rho(\Gamma)$$
$$\mathbb{E}[< \partial_t u_p, v > + \mathcal{B}(u_p, v)] = \mathbb{E}[\int_D f\, v], \qquad \forall v \in H_0^1(D) \otimes \mathcal{P}_p(\Gamma)$$

which gives the Galerkin orthogonality

$$\mathbb{E}[< \partial_t e_p, v > + \mathcal{B}(e_p, v)] = 0,\ \forall v \in H_0^1(D) \otimes \mathcal{P}_p(\Gamma). \tag{21}$$

Now consider a function $w \in L^2(0,T;\mathcal{P}_p(\Gamma)\otimes H_0^1(D))$ with $\partial_t w \in L^2(0,T;\mathcal{P}_p(\Gamma)\otimes H^{-1}(D))$ and such that $u(0,\cdot) = u_p(0,\cdot) = w(0,\cdot)$.

Then we have

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}[\|(u_p - w)(T,\cdot)\|^2_{L^2(D)}] + a_{\min}\mathbb{E}[\|u_p - w\|^2_{L^2(0,T;H_0^1(D))}] \\
&\le \int_0^T \mathbb{E}[< \partial_t(u_p - w), u_p - w > + \mathcal{B}(u_p - w, u_p - w)]dt \\
&\le \int_0^T \underbrace{\mathbb{E}[< \partial_t(u_p - u), u_p - w > + \mathcal{B}(u_p - u, u_p - w)]}_{=0 \text{ by Galerkin orthogonality}} dt \\
&+ \int_0^T \mathbb{E}[< \partial_t(u - w), u_p - w > + \mathcal{B}(u - w, u_p - w)]dt.
\end{aligned}$$

Now estimate the terms in the above integral by

$$|\mathbb{E}[< \partial_t(u - w), u_p - w >]| \le \mathbb{E}[\|\partial_t(u - w)\|_{H^{-1}(D)}\|u_p - w\|_{H_0^1(D)}]$$

and

$$|\mathbb{E}[\mathcal{B}(u - w, u_p - w)]| \le a_{\max}\mathbb{E}[\|u - w\|_{H_0^1(D)}\|u_p - w\|_{H_0^1(D)}].$$

Then apply the inequality $\alpha\beta \le \frac{\alpha^2}{2\epsilon} + \frac{\epsilon\beta^2}{2}$ and arrive at

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}[\|(u_p - w)(T,\cdot)\|^2_{L^2(D)}] + \frac{a_{\min}}{2}\mathbb{E}[\|u_p - w\|^2_{L^2(0,T;H_0^1(D))}] \\
&\le \frac{1}{a_{\min}}\mathbb{E}[\|\partial_t(u - w)\|^2_{L^2(0,T;H^{-1}(D))}] + \frac{a^2_{\max}}{a_{\min}}\mathbb{E}[\|u - w\|^2_{L^2(0,T;H_0^1(D))}].
\end{aligned}$$

Finally, summing on both sides of the last inequality the terms

$$\frac{1}{2}\mathbb{E}[\|(u - w)(T,\cdot)\|^2_{L^2(D)}] + \frac{a_{\min}}{2}\mathbb{E}[\|u - w\|^2_{L^2(0,T;H_0^1(D))}],$$

and applying the triangular inequality yields

$$\begin{aligned}
&\frac{1}{4}\mathbb{E}[\|(u_p - u)(T,\cdot)\|^2_{L^2(D)}] + \frac{a_{\min}}{4}\mathbb{E}[\|u_p - u\|^2_{L^2(0,T;H_0^1(D))}] \\
&\le \frac{1}{2}\mathbb{E}[\|(u - w)(T,\cdot)\|^2_{L^2(D)}] + \frac{1}{a_{\min}}\mathbb{E}[\|\partial_t(u - w)\|^2_{L^2(0,T;H^{-1}(D))}] \\
&+ \left(\frac{a^2_{\max}}{a_{\min}} + \frac{a_{\min}}{2}\right)\mathbb{E}[\|u - w\|^2_{L^2(0,T;H_0^1(D))}],
\end{aligned}$$

which is what we wanted to prove. □

The last Theorem is useful in the study of the convergence of Stochastic Galerkin approximations. Indeed, we reduce the problem of estimating the size of the error $e_p = u - u_p$ to that of estimating the *best approximation error* in $L^2_\rho(\Gamma)$. Observe, however, that being ρ bounded, the best approximation error can be equivalently stated, up to a multiplicative constant, in $L^2(\Gamma)$, with respect to the Lebesgue measure (see next section) and therefore no longer weighted by the, in general non uniform, probability density ρ. This allows us to apply standard approximation results for functions in $L^2(\Gamma)$.

5 Convergence analysis for Stochastic Galerkin

Observe that from (20) we can reduce the general case of a bounded probability density ρ to the case of uniform, independent random variables. Indeed, it is enough for this purpose to apply the inequality $\int_\Gamma |Z|(\mathbf{y})\rho(\mathbf{y})d\mathbf{y} \leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)} \int_\Gamma |Z|(\mathbf{y})\hat{\rho}(\mathbf{y})d\mathbf{y}$ with $\hat{\rho} = \frac{1}{|\Gamma|}$ in all the expected values appearing in (20). Thus, we only need to consider the case ρ constant when proving the results, just as it was done in [1].

Following [13, 1], we use the Legendre polynomials to estimate best $L^2(\Gamma)$ approximation properties of the polynomial spaces $\mathcal{P}_p(\Gamma)$ introduced in examples 1 and 2, respectively. These estimates, combined with the Stochastic Galerkin optimality (20) will yield error estimates for the $p\times h-$version of the SGFEM. Without loss of generality we now assume that $\Gamma = [-1,1]^N$, for instance after making a linear change of variables as in Section 3. Now we consider an orthogonal polynomial basis for $L^2([-1,1])$, namely the Legendre polynomials,

$$\phi_n(s) \equiv \frac{1}{2^n n!}\frac{d^n}{ds^n}((s^2-1)^n), \quad n = 0, 1, \ldots,$$

and a representation of the polynomial subspace $\mathcal{P}_p(\Gamma)$ in terms of Legendre polynomials and a set of multi-indices $\mathcal{I}_{\mathcal{P}_p(\Gamma)}$ such that when we let the multi-index $\mathbf{i}$ vary over $\mathcal{I}_{\mathcal{P}_p(\Gamma)}$ the corresponding multivariate Legendre polynomials span the subspace $\mathcal{P}_p(\Gamma)$, namely

$$\mathcal{P}_p(\Gamma) = \text{span}\{\phi_{\mathbf{i}}, \ \ \mathbf{i} \in \mathcal{I}_{\mathcal{P}_p(\Gamma)}\}, \qquad \phi_{\mathbf{i}}(\mathbf{y}) = \prod_{n=1}^{N} \phi_{i_n}(y_n).$$

Indeed, we work here with two cases,

anisotropic tensor polynomials: $\mathcal{I}_{\mathcal{P}_{\mathbf{p}}(\Gamma)} = \{\mathbf{i} \in \mathbb{N}_+^N : i_n \leq p_n, n = 1, \ldots, N\}$,

total degree polynomials: $\mathcal{I}_{\mathcal{P}_p(\Gamma)} = \{\mathbf{i} \in \mathbb{N}_+^N : \|\mathbf{i}\|_{\ell^1} \leq p\}$.

Since the polynomials $\phi_{\mathbf{i}}$ are orthogonal in $L^2([-1,1]^N)$ we can introduce a suitable projection of the exact solution u, namely w as follows. We choose

$w \in \mathcal{P}_p(\Gamma) \otimes L^2(0,T;H_0^1(D))$ such that at each time of $[0,T]$ it is the $L^2(\Gamma)$-projection of $u(t)$ over $\mathcal{P}_p(\Gamma)$, i.e.

$$\int_\Gamma w(t)v = \int_\Gamma u(t)v, \text{ for all } v \in \mathcal{P}_p(\Gamma) \text{ and a.e. on } [0,T].$$

With this choice of w we have, by Theorem 1,

$$\begin{aligned}
&\mathbb{E}[\|(u_p-u)(T,\cdot)\|^2_{L^2(D)}] + a_{\min}\mathbb{E}[\|u_p-u\|^2_{L^2(0,T;H_0^1(D))}] \\
&\le C\|\rho\|_{L^\infty(\Gamma)}\Big(\|(u-w)(T)\|^2_{L^2(\Gamma)\otimes L^2(D)} + \|u-w\|^2_{L^2(\Gamma)\otimes L^2(0,T;H_0^1(D))} \\
&+ \|\partial_t(u-w)\|^2_{L^2(\Gamma)\otimes L^2(0,T;H_0^{-1}(D))}\Big) \\
&\le C\|\rho\|_{L^\infty(\Gamma)} \sum_{\mathbf{i}\in\mathbb{N}^N\setminus\mathcal{I}_{\mathcal{P}_p(\Gamma)}} \frac{\|d_\mathbf{i}(T)\|^2_{L^2(D)} + \|d_\mathbf{i}\|^2_{L^2(0,T;H_0^1(D))} + \|\partial_t d_\mathbf{i}\|^2_{L^2(0,T;H^{-1}(D))}}{\|\phi_\mathbf{i}\|^2_{L^2(\Gamma)}}
\end{aligned}$$

with the function valued Fourier coefficients

$$d_\mathbf{i} = \int_\Gamma u(\mathbf{y},\cdot)\phi_\mathbf{i}(\mathbf{y})d\mathbf{y}. \tag{22}$$

Indeed, similar arguments as in [1] show that the following estimate for the size of the Fourier coefficients of the exact solution, u, holds.

Lemma 2 (Fourier coefficients estimate) *Let $\mathbf{i} \in \mathbb{N}^N$, and $d_\mathbf{i}$ defined in (22) with $u(\mathbf{y},\cdot)$ being the solution to (7). Then there exists a constant $C = C(f, a_{min}, a_{max}, u_0, C_D) > 0$ not depending on $\mathbf{i}$ such that*

$$\begin{aligned}
&\frac{\|d_\mathbf{i}(T)\|^2_{L^2(D)} + \|d_\mathbf{i}\|^2_{L^2(0,T;H_0^1(D))} + \|\partial_t d_\mathbf{i}\|^2_{L^2(0,T;H^{-1}(D))}}{\|\phi_\mathbf{i}\|^2_{L^2(\Gamma)}} \\
&\le C|\Gamma|(3\pi)^N \left\{\prod_{n=1}^N \left(\sqrt{1-e^{-2g_n}} + \mathcal{O}\left(\frac{1}{i_n^{1/3}}\right)\right) e^{-g_n i_n}\right\}^2 \\
&\le \tilde{C}|\Gamma| e^{-2\sum_{n=1}^N g_n i_n}
\end{aligned}$$

with $g_n \equiv \log(1 + r_n + \sqrt{r_n^2 + 2r_n})$ and r_n defined in (9) and

$$\tilde{C} = C(3\pi)^N \prod_{n=1}^N \Big(\sqrt{1-e^{-2g_n}} + \mathcal{O}(1)\Big)^2. \tag{23}$$

5.1 Convergence analysis for anisotropic tensor product approximations

We recall that for an anisotropic tensor product polynomial approximation the index set is defined as $\mathcal{I}_{\mathcal{P}_\mathbf{p}(\Gamma)} = \{\mathbf{i} \in \mathbb{N}_+^N : i_n \le p_n, n = 1,\dots,N\}$. The main

result of this section, namely the exponential convergence with respect to the multi-index $\mathbf{p} = (p_1, \ldots, p_n)$ as in [13], follows from the above lemma. Here we include the final convergence results for the approximation of the solution to (4) with the anisotropic p-version.

Theorem 2 (Convergence w.r.t. the multi-index p) *With the same assumptions as in Theorem 1 we have*

$$\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}]+a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}] \leq \ |\Gamma|\|\rho\|_{L^\infty(\Gamma)}\tilde{C}\sum_{n=1}^{N} e^{-2g_n(p_n+1)}$$

with $g_n > 0$ defined in (9) and $\tilde{C} > 0$ independent of p_n and ρ defined in (23).

Proof. To obtain the result, combine the results from Theorem 1 and Lemma 2. □

Recall now that the number of degrees of freedom in the tensor approximation is $\eta = \prod_{n=1}^{N}(1+p_n) \leq e^{\sum_{n=1}^{N} p_n}$. This estimate combined with Theorem 2 yields

Theorem 3 (Algebraic convergence w.r.t. to η) *Let p be a positive integer and choose the polynomial degree in the $n-th$ direction, p_n, to be the smallest integer such that $p_n \geq p\frac{g_{\min}}{g_n}$, $n = 1, \ldots, N$. Then we have*

$$\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}] + a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}] \leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)}\tilde{C}\left(\sum_{n=1}^{N} e^{-2g_n}\right)\eta^{-\frac{2}{\sum_{n=1}^{N} 1/g_n}}$$

5.2 Convergence analysis for total degree polynomial approximations

Again, we recall that for the total degree polynomial space the index set is $\mathcal{I}_{\mathcal{P}_p(\Gamma)} = \{\mathbf{i} \in \mathbb{N}^N_+ : \|\mathbf{i}\|_{\ell^1} \leq p\}$. Similarly as in Section 5.1, using the optimality of the Stochastic Galerkin approximation proved in Theorem 1 and the estimate of the Fourier coefficients in Lemma 2 we obtain

Theorem 4 (Convergence w.r.t. to the total degree p) *We have*

$$\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}] + a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}] \leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)}\tilde{C}\max\left(\frac{N-1}{g_{\min}}, 2+p\right)^{N-1}\frac{e^{-2g_{\min}(1+p)}}{1-e^{-g_{\min}}}. \tag{24}$$

If in addition, N is sufficiently small such that

$$\beta(N) \equiv \begin{cases} 2g_{\min} & \text{for } N = 1 \\ 2g_{\min} - 1 - \log(N-1) & \text{for } N > 1 \end{cases} \tag{25}$$

satisfies $\beta(N) > 0$, then we have the exponential convergence

$$\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}]+a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}] \leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)}\tilde{C}\frac{e^{-\beta(N)(p+1)}}{1-e^{-\beta(N)}}. \tag{26}$$

Proof. Let $\mathcal{E}^2 = \frac{1}{\|\rho\|_{L^\infty(\Gamma)}}\left\{\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}] + a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}]\right\}$. We have

$$\begin{aligned}
\mathcal{E}^2 &\leq C\sum_{|\mathbf{i}|>p}\frac{\|d_\mathbf{i}(T)\|^2_{L^2(D)} + \|d_\mathbf{i}\|^2_{L^2(0,T;H^1_0(D))} + \|\partial_t d_\mathbf{i}\|^2_{L^2(0,T;H^{-1}(D))}}{\|\phi_\mathbf{i}\|^2_{L^2(\Gamma)}} \\
&\leq |\Gamma|\tilde{C}\sum_{|\mathbf{i}|>p} e^{-2\sum_{n=1}^N g_n i_n} \\
&\leq |\Gamma|\tilde{C}\sum_{|\mathbf{i}|>p} e^{-2g_{\min}\sum_{n=1}^N i_n} \\
&\leq |\Gamma|\tilde{C}\sum_{s=p+1}^{+\infty} e^{-2g_{\min}s}\binom{N-1+s}{N-1}
\end{aligned}$$

Observe that in the last inequality we have used that

$$\#\{\mathbf{i}\in\mathbb{N}_+^N : |\mathbf{i}| = s\} = \binom{N-1+s}{N-1}$$

and that we can further bound, using that $\log(1+x)\leq x$, for $0\leq x$,

$$\begin{aligned}
\binom{N-1+s}{N-1} &= \prod_{n=1}^{N-1}\left(1+\frac{s}{n}\right) \\
&\leq \min\left(e^{s\sum_{n=1}^{N-1}\frac{1}{n}}, (1+s)^{N-1}\right).
\end{aligned}$$

Now we employ the inequality $\sum_{n=2}^N \frac{1}{n} \leq \int_1^N \frac{dx}{x} = \log(N)$ to arrive at

$$\binom{N-1+s}{N-1} \leq \begin{cases} 1, \text{ for } N=1, \\ \min\left(e^{s(1+\log(N-1))}, (1+s)^{N-1}\right), \text{ for } N>1, \end{cases}$$

and therefore, for $N>1$,

$$\mathcal{E}^2 \leq |\Gamma|\tilde{C}\sum_{s=p+1}^{+\infty} e^{-2g_{\min}s}\min\left(e^{s(1+\log(N-1))}, (1+s)^{N-1}\right).$$

On the other hand, the function $f(s) = (1+s)^{N-1}e^{-g_{min}s}$ has a maximum in $\bar{s} = (N-1)/g_{min} - 1$ and $\forall s \geq p+1$ it holds

$$\begin{aligned}
&\text{if } \bar{s}\geq p+1 \qquad && f(s)\leq f(\bar{s}) \leq \left(\frac{N-1}{g_{min}}\right)^{N-1} e^{-g_{min}(p+1)}, \\
&\text{if } \bar{s} < p+1 \qquad && f(s)\leq f(p+1) = (p+2)^{N-1}e^{-g_{min}(p+1)}.
\end{aligned}$$

Therefore,

$$f(s) \leq \max\left(\frac{N-1}{g_{\min}}, 2+p\right)^{N-1} e^{-g_{\min}(1+p)}, \text{ for all } s \geq p+1$$

and we have, for $N > 1$,

$$\begin{aligned}\mathcal{E}^2 &\leq |\Gamma|\tilde{C} \min\left\{\sum_{s=p+1}^{+\infty} e^{s((1-2g_{\min})+\log(N-1))}, \sum_{s=p+1}^{+\infty} e^{-g_{\min}s} f(s)\right\} \\ &\leq \begin{cases} \frac{|\Gamma|\tilde{C}}{1-e^{-\beta(N)}} e^{-\beta(N)(p+1)}, & \text{if } \beta(N) > 0 \\ \frac{|\Gamma|\tilde{C}}{1-e^{-g_{\min}}} \max\left(\frac{N-1}{g_{\min}}, 2+p\right)^{N-1} e^{-2g_{\min}(1+p)} & \forall p > 0,\ N > 1. \end{cases}\end{aligned}$$

□

Remark 2 *The upper bounds in the previous theorem are difficult to improve. For instance, by including more terms in the expansion of* $\log(1+x)$ *one eventually obtains an upper bound of the form*

$$|\Gamma|\tilde{C} \sum_{s=p+1}^{+\infty} e^{-2g_{\min}s + s(1+log(N-1)) + f(s)}$$

for a function f *that does not depend on* N*. This upper bound still exhibits a similar deterioration with respect to the input dimension,* N*, as* (26)*.*

Now we present the corresponding result with respect to the number of degrees of freedom in the total degree polynomial space. Indeed, we have in this case the following upper and lower bounds for the number η of degrees of freedom in $\mathcal{P}_p(\Gamma)$,

$$\begin{aligned}\max\left(Np, \frac{p^N}{N!}\right) \leq \eta &= \prod_{n=1}^{N}\left(1+\frac{p}{n}\right) \\ &\leq e^{p\sum_{j=1}^{N}\frac{1}{n}} \\ &\leq e^{p(1+\log(N))}\end{aligned} \tag{27}$$

and therefore, the combination of (26) from Theorem 4 and (27) yields

Theorem 5 (Convergence w.r.t. to η) *We have*

$$\begin{aligned}&\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}] + a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}] \\ &\leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)}\tilde{C}\frac{e^{-2g_{\min}}}{1-e^{-g_{\min}}} \max\left(\frac{N-1}{g_{\min}}, 2+\min\left\{\frac{\eta}{N}, \eta^{1/N}(N!)^{1/N}\right\}\right)^{N-1} \eta^{-\frac{2g_{\min}}{1+\log(N)}}.\end{aligned} \tag{28}$$

If in addition, $\beta(N)$ defined in (25) *is positive, then we have the algebraic convergence*

$$\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}] + a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}] \\ \leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)}\frac{\tilde{C}e^{-\beta(N)}}{1-e^{-\beta(N)}}\eta^{-\frac{\beta(N)}{1+\log(N)}}.$$

Remark 3 *Observe that* (28) *is not optimal, specially because the lower bound in* (27) *is not tight enough. For instance, if we use the alternative lower bound*

$$\eta \geq \exp\left(\sum_{n=1}^{N}\frac{p}{n} - \frac{1}{2}\sum_{n=1}^{N}\frac{p^2}{n^2}\right),$$

for small values of the degree p, say,

$$p \leq \frac{\log(N+1)}{2\sum_{j=1}^{+\infty}\frac{1}{j^2}}$$

then we have, in terms of the corresponding number of degrees of freedom, $\eta(p)$, the estimate

$$\mathbb{E}[\|(u-u_p)(T)]\|^2_{L^2(D)}] + a_{min}\,\mathbb{E}[\|u-u_p\|^2_{L^2(0,T;H^1_0(D))}] \\ \leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)}\tilde{C}\frac{e^{-2g_{\min}}}{1-e^{-g_{\min}}}\max\left(\frac{N-1}{g_{\min}}, 2(1+\frac{\log(\eta)}{\log(N+1)})\right)^{N-1}\eta^{-\frac{2g_{\min}}{1+\log(N)}}.$$

6 Stochastic Collocation

Stochastic Collocation has gained much attention recently from the computational community, see for instance the works [3, 10, 23, 20, 32]. This technique can be based on either full or sparse tensor product approximation spaces, as we describe in what follows.

6.1 Full tensor product interpolation

In this section we briefly recall interpolation based on Lagrange polynomials, see Section 2 in [3] and as in Section 5 we assume $\Gamma = [-1,1]^N$. We first introduce a non negative index $i \geq 1$ and then, for each value of i, let $\{y^i_1, \ldots, y^i_{m_i}\} \subset [-1,1]$ be a sequence of abscissas for Lagrange interpolation on $[-1,1]$.

Let we denote by $W(D)$ a Banach space of functions where the solution of $u(\mathbf{y},\cdot)$ takes value. For instance, in our case $W(D)$ contains functions from $L^2(0,T;H^1_0(D))$ whose time derivatives take values in $L^2(0,T;H^{-1}(D))$.

For $u \in C^0(\Gamma^1;W(D))$ and $N=1$ we introduce a sequence of one-dimensional Lagrange interpolation operators $\mathscr{U}^i : C^0(\Gamma^1;W(D)) \to V_{m_i}(\Gamma^1;W(D))$

$$\mathscr{U}^i(u)(y) = \sum_{j=1}^{m_i} u(y^i_j)\, l^i_j(y), \quad \forall u \in C^0(\Gamma^1;W(D)), \tag{29}$$

where $l_j^i \in \mathcal{P}_{m_i-1}(\Gamma^1)$ are the Lagrange polynomials of degree m_i-1, i.e. $l_j^i(y) = \prod_{\substack{k=1\\k\neq j}}^{m_i} \frac{(y-y_k^i)}{(y_j^i-y_k^i)}$, and

$$V_m(\Gamma^1; W(D)) = \left\{ v \in C^0(\Gamma^1; W(D)) \,:\, v(\mathbf{y},\mathbf{x}) = \sum_{k=1}^{m} \widetilde{v}_k(\mathbf{x}) l_k(\mathbf{y}),\ \{\widetilde{v}_k\}_{k=1}^m \in W(D) \right\}.$$

Formula (29) reproduces exactly all polynomials of degree less than m_i. Now, in the multivariate case $N > 1$, for each $u \in C^0(\Gamma^N; W(D))$ and the multi-index $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{N}_+^N$ we recall the full tensor product interpolation formula,

$$\begin{aligned} \hat{\Pi}_{\mathbf{i}}^N u(\mathbf{y}) &= \left(\mathscr{U}^{i_1} \otimes \cdots \otimes \mathscr{U}^{i_N}\right)(u)(\mathbf{y}) \\ &= \sum_{j_1=1}^{m_{i_1}} \cdots \sum_{j_N=1}^{m_{i_N}} u\left(y_{j_1}^{i_1}, \ldots, y_{j_N}^{i_N}\right) \left(l_{j_1}^{i_1} \otimes \cdots \otimes l_{j_N}^{i_N}\right)(\mathbf{y}). \end{aligned} \tag{30}$$

Clearly, the above product needs $\prod_{n=1}^N m_{i_n}$ function evaluations. These formulas will also be used as the building blocks for the Smolyak method, described next.

6.2 Stochastic Collocation based on isotropic sparse grids

Here we follow closely the work [4] and describe the Smolyak *isotropic* formulas $\mathscr{A}(\mathrm{w}, N)$. The Smolyak formulas are just linear combinations of tensor product formulas (30) where the indices are chosen such that only tensor products with a relatively small number of points are used. With $\mathscr{U}^0 = 0$ and for $i \in \mathbb{N}_+$ define

$$\Delta^i := \mathscr{U}^i - \mathscr{U}^{i-1}. \tag{31}$$

Moreover, given a positive integer $\mathrm{w} \in \mathbb{N}_+$, hereafter called the *level*, and a multi-index $\mathbf{i} \in \mathbb{N}_+^N$, the isotropic Smolyak formula is given by

$$\mathscr{A}(\mathrm{w}, N) = \sum_{|\mathbf{i}| \leq \mathrm{w}+N} \left(\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_N}\right). \tag{32}$$

Equivalently, formula (32) can be written as (see [31])

$$\mathscr{A}(\mathrm{w}, N) = \sum_{\mathrm{w}+1 \leq |\mathbf{i}| \leq \mathrm{w}+N} (-1)^{\mathrm{w}+N-|\mathbf{i}|} \binom{N-1}{\mathrm{w}+N-|\mathbf{i}|} \cdot \left(\mathscr{U}^{i_1} \otimes \cdots \otimes \mathscr{U}^{i_N}\right). \tag{33}$$

To compute $\mathscr{A}(\mathrm{w}, N)(u)$, one only needs to know function values on the "sparse grid"

$$\mathscr{H}(\mathrm{w}, N) = \bigcup_{\mathrm{w}+1 \leq |\mathbf{i}| \leq \mathrm{w}+N} \left(\vartheta^{i_1} \times \cdots \times \vartheta^{i_N}\right) \subset [-1,1]^N, \tag{34}$$

where $\vartheta^i = \left\{y_1^i, \ldots, y_{m_i}^i\right\} \subset [-1,1]$ denotes the set of abscissas used by $\mathscr{U}^i$. If the sets are nested, i.e. $\vartheta^i \subset \vartheta^{i+1}$, then $\mathscr{H}(\mathrm{w}, N) \subset \mathscr{H}(\mathrm{w}+1, N)$ and

$$\mathscr{H}(\mathrm{w}, N) = \bigcup_{|\mathbf{i}| = \mathrm{w}+N} \left(\vartheta^{i_1} \times \cdots \times \vartheta^{i_N}\right). \tag{35}$$

The Smolyak formula is actually interpolatory whenever nested points are used. This result has been proved in [4, Proposition 6 on page 277].

By comparing (35) and (34), we observe that the Smolyak approximation that employs nested points requires less function evaluations than the corresponding formula with non nested points.

Clenshaw-Curtis abscissas. A popular choice for abscissas in the construction of the Smolyak formula are the Clenshaw-Curtis ones [6]. These abscissas are the extrema of Chebyshev polynomials and, for any choice of $m_i > 1$, are given by

$$y_j^i = -\cos\left(\frac{\pi(j-1)}{m_i - 1}\right), \quad j = 1, \ldots, m_i. \tag{36}$$

In addition, one sets $y_1^i = 0$ if $m_i = 1$ and lets the number of abscissas m_i in each level to grow according to the following formula

$$m_1 = 1 \quad \text{and} \quad m_i = 2^{i-1} + 1, \text{ for } i > 1. \tag{37}$$

With this particular choice, one obtains nested sets of abscissas, i.e., $\vartheta^i \subset \vartheta^{i+1}$ and thereby $\mathscr{H}(\mathrm{w}, N) \subset \mathscr{H}(\mathrm{w}+1, N)$. It is important to choose $m_1 = 1$ if we are interested in optimal approximation in relatively large N, because in all other cases the number of points used by $\mathscr{A}(\mathrm{w}, N)$ increases too fast with N.

Remark 4 (Anisotropic Smolyak) *The work [22] proposed and analyzed a novel* anisotropic *sparse grid stochastic collocation method that is based on a weighted version of the Smolyak formula* (32)*, namely one penalizes the use of approximation levels on input random variables which have little influence in the solution. In other words, given a vector with positive weights* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N) \in \mathbb{R}_+^N$*, the anisotropic formula is*

$$\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N) = \sum_{\sum_{n=1}^N (i_n - 1)\alpha_n \leq \mathrm{w}} \left(\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_N}\right) \tag{38}$$

for each level value $\mathrm{w} \in \mathbb{N}_+$*. The work [22] also developed a procedure for choosing the anisotropy of the sparse grid (i.e. the weight vector* $\boldsymbol{\alpha}$*, based on either* a priori *or* a posteriori *estimates and showed both theoretically and numerically the effectiveness of these methods for several problems. This approach is particularly attractive in the case of truncated expansions of random fields, since the anisotropy can be tuned to the decay properties of the expansion.*

7 Convergence analysis for Stochastic Collocation

Here we recall convergence results from previous works to motivate the use of Stochastic Collocation methods.

7.1 Convergence analysis for full tensor grids

An isotropic full tensor product interpolation converges roughly like $C(g_{min}, N)e^{-g_{min}p}$, see [3], where p is the order of the polynomial space. Since the number of collocation points relates to p in this case as $\eta = (1+p)^N$ then $\log(\eta) = N\log(1+p) \leq Np$ and with respect to η the convergence rate can be bounded as $C(g_{min}, N)\eta^{-g_{min}/N}$, with $g_{min} > 0$ being a problem dependent constant. The slowdown effect that the dimension N has on the last convergence is known as the curse of dimensionality and it is the reason for not using isotropic full tensor interpolation for large values of N.

On the other hand, the behavior of anisotropic approximations is a bit better, and tends to follow the one corresponding to Stochastic Galerkin based on anisotropic full tensor products, see Theorem 3. Actually, it was shown in [3] that in some cases, polynomial approximations computed by the full tensor grid Stochastic Collocation method coincide with the ones obtained by Stochastic Galerkin projecting onto the same tensor polynomial space.

This applies in particular to problem (1) with diffusion coefficient (6) provided the density ρ of the random vector $[Y_1, \ldots, Y_N]$ factorizes as $\rho(\mathbf{y}) = \prod_{n=1}^{N} \rho_n(y_n)$ (meaning that the random variables are independent) and we choose Gauss abscissas in each direction with respect to the weight ρ_n. If, in addition, we consider an anisotropic tensor grid in which the number of knots in the n-th direction is chosen as $i_n = \frac{g_{min}}{g_n}p + 1$, then the rate of convergence is the one stated in Theorem 3.

We point out, however, that in more general cases of nonlinear problems or non-linear expansions of the input random fields, the Stochastic Collocation solution does not coincide with the Stochastic Galerkin one, the main advantage of the first approach being that it always produces a set of uncoupled equations, while preserving roughly the same accuracy of the second one.

7.2 Convergence analysis for isotropic sparse grids

Whenever the number of input random variables, N, is relatively large, the isotropic sparse grid approximation seems to be better suited than a full tensor one.

In [23] we have derived a general convergence result for the isotropic Smolyak approximation of Banach-valued functions $u \in C^0(\Gamma; W)$, where W is an arbitrary Banach space, under the assumption that u is analytic with respect to each input variable $y_n \in \Gamma_n$. This assumption holds in our case, as we showed in Section 3. The following result can therefore be proved with minimal changes of the proof in [23]:

Theorem 6 *Let* w *be a positive integer and* η(w) *the total number of points in the isotropic sparse grid based on Clenshaw-Curtis abscissas. With the same*

assumptions as in Theorem 1 we have

$$\mathbb{E}[\|(u-\mathscr{A}(\mathrm{w},N)u)(T)]\|^2_{L^2(D)}] + a_{min}\,\mathbb{E}[\|u-\mathscr{A}(\mathrm{w},N)u\|^2_{L^2(0,T;H^1_0(D))}] \\ \leq |\Gamma|\|\rho\|_{L^\infty(\Gamma)} C(\tilde{g}_{min}, a_{min}, a_{max}, f, C_D, N)\eta(\mathrm{w})^{-\frac{\tilde{g}_{min}}{1+log(2N)}} \quad (39)$$

where $\tilde{g}_n = \frac{1}{2}\log\left(2r_n + \sqrt{1+4r_n^2}\right)$, $\tilde{g}_{min} = \min_n g_n$ *and* r_n *defined as in Lemma 9.*

Observe that the previous result indicates at least *algebraic convergence* with respect to the number of collocation points η and we have a similar degradation of the exponent with respect to the dimension N as in Theorem 5 for the total degree polynomial - Stochastic Galerkin approximation. However, differently than (28) there are no polynomial terms in η multiplying the leading term in the estimate.

For large values of w we have a related subexponential convergence result, see [23]. Finally, we point out that the convergence results hold also if one chooses Gaussian abscissas instead of Clenshaw-Curtis ones (see [23]).

8 Numerical Example

In this numerical example we compare the performance of different numerical approximations. We consider problem (1) with a stochastic diffusion coefficient, zero forcing term f and zero initial condition u_0. The physical domain is the unit square $[0,1]^2$. Homogeneous Dirichlet boundary conditions are imposed on the bottom, right and top edges, while an incoming unitary flux is imposed on the left edge, namely $a\partial_n u = 1$. The domain and spatial mesh employed in the simulations are shown in Figure 1. All the approximations use the same time and space discretizations. They only differ in the treatment of the stochastic discretization. The random diffusion coefficient varies only in the vertical direction and has the following form

$$a(\omega,x,y) = a_0 + \sigma\sqrt{\lambda_0}Y_0(\omega) + \sum_{i=1}^{n_f}\sigma\sqrt{\lambda_i}\left[Y_i(\omega)\cos(i\pi y) + Y_{n_f+i}(\omega)\sin(i\pi y)\right] \quad (40)$$

with

$$\lambda_0 = \frac{\sqrt{\pi}L_c}{2}, \qquad \lambda_i = \sqrt{\pi}L_c e^{-\frac{(i\pi L_c)^2}{4}}, \;\; i = 1\ldots,n_f$$

and $Y_0,\ldots,Y_{2n_f}$ uncorrelated zero mean and unit variance random variables. Expansion (40) approximates a stationary random field with covariance function

$$\mathrm{Cov}[a](y_1,y_2) \approx \sigma^2\exp\left\{-\frac{(y_1-y_2)^2}{L_c^2}\right\}.$$

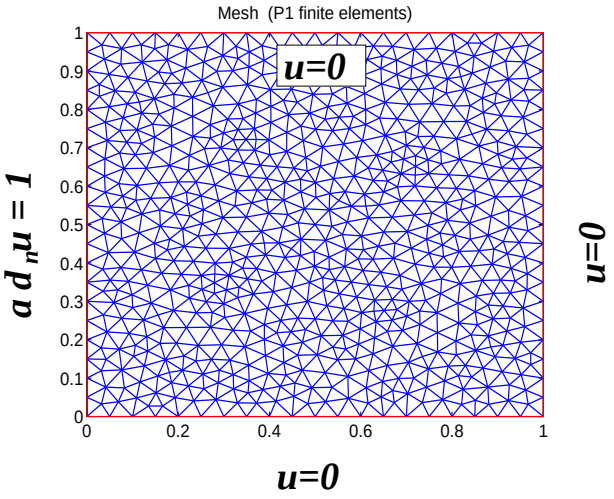

Figure 1: Computational mesh and boundary conditions for Example 2

The parameter L_c represents a physical correlation length and σ^2 the variance of the random field.

In the numerical simulations we have used the following values: $a_0 = 1$, $L_c = 0.25$, $\sigma = 0.15$, $n_f = 3$ (corresponding to $N = 7$ random variables) and we have assumed the random variables $Y_0, \ldots, Y_{2n_f}$ independent and uniformly distributed in the interval $[-\sqrt{3}, \sqrt{3}]$.

In particular the choice $\sigma = 0.15$ guarantees that the random field is strictly positive, even in the limit $n_f \to \infty$. On the other hand, by choosing $n_f = 3$ we represent 95% of the variance of the limit field for $n_f \to \infty$. Figure 2 shows 4 random realizations of the truncated random field (40).

The deterministic solver employs continuous piecewise linear finite elements in space and the implicit Euler method in time. We have run 40 uniform time steps in the interval $[0, T = 0.1]$ and focused on the quantity of interest

$$\psi(\omega) = \int_D u(\omega, T, x) dx.$$

Figures 3 and 4 show the mean and the standard deviation of the solution at the final time $T = 0.1$, while Figure 5 shows a sample histogram of the quantity of interest ψ, obtained by Monte Carlo sampling. We see in Figure 5 that the distribution of ψ slightly deviates from a Gaussian one.

On this example, we have compared several techniques in the computation of the mean value of the quantity of interest, i.e. $\mathbb{E}[\psi]$. The results are summarized in Figure 7, which shows convergence plots obtained by comparing the solution computed by the different methods, with an overkilling solution computed with the level 5 isotropic Smolyak method (25978 collocation points) using the same spatial and temporal grid. The errors shown are relative to the value of the

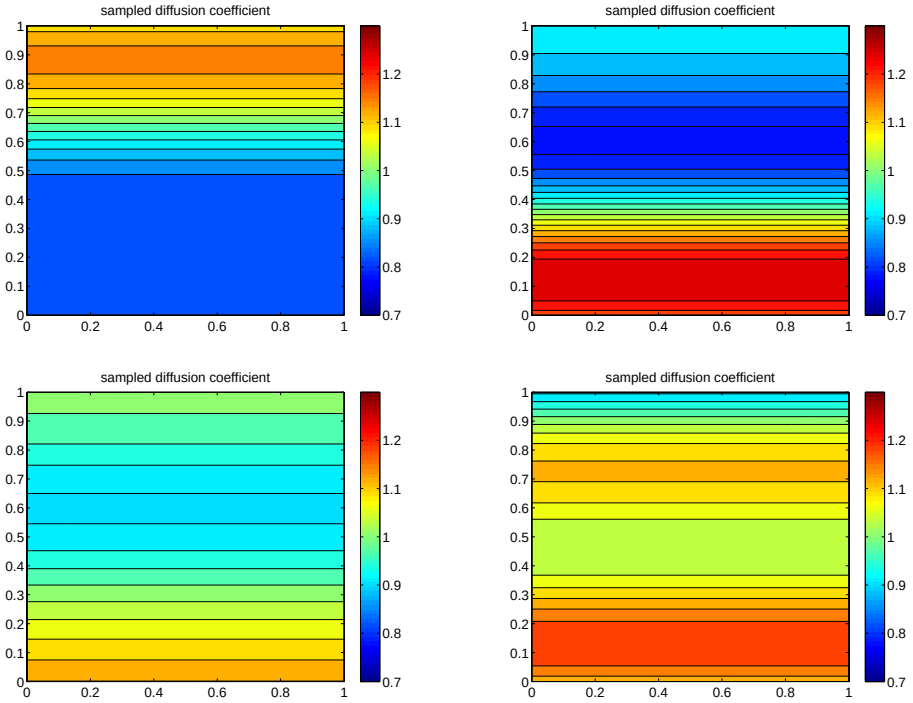

Figure 2: Realizations of the diffusivity coefficient.

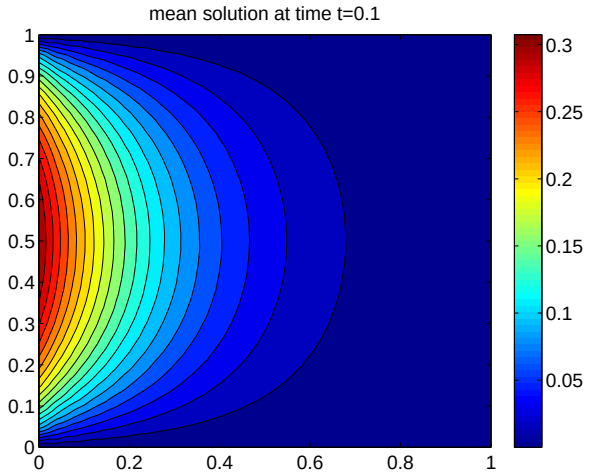

Figure 3: Mean of the solution at time 0.1.

overkilling solution. All simulations have been run in Matlab using the PDE toolbox.

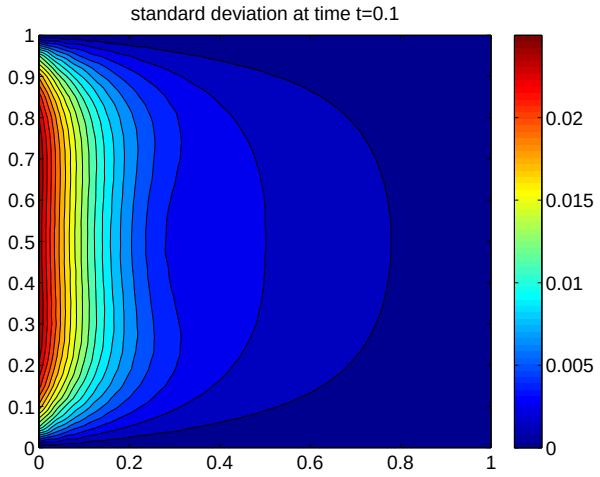

Figure 4: Standard deviation of the solution at time 0.1.

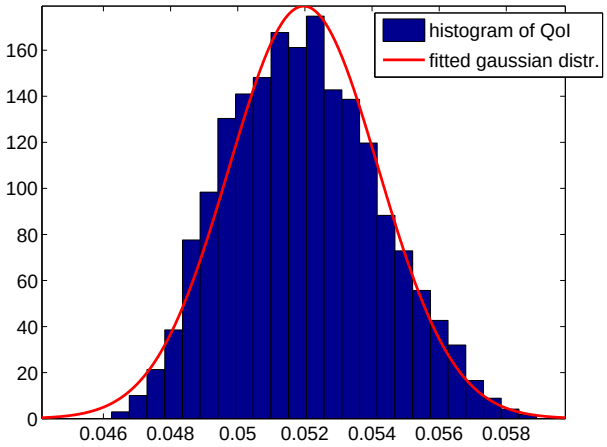

Figure 5: Histogram of the quantity of interest.

Anisotropic tensor product Stochastic Galerkin. The first methodology considered is the Stochastic Galerkin method based on an anisotropic tensor product polynomial space. Since the problem is linear and the diffusion coefficient depends linearly on the random variables, a proper choice of basis functions (double orthogonal polynomials, see [1]) allows us to decouple the global system and reduce it to a sequence of uncoupled deterministic problems. It can be shown that in such a case, the Stochastic Galerkin method coincides with a Stochastic Collocation on a proper tensor grid of Gauss points (see [3]).

The polynomial degrees in each direction have been chosen as suggested in

	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	η
$p=1$	0	1	0	0	1	0	0	4
$p=2$	1	2	1	1	2	1	1	288
$p=3$	2	3	2	2	3	2	2	3888
$p=4$	3	4	3	2	4	3	2	14400

Table 1: Polynomial degree used in the different directions, for $p = 1, 2, 3, 4$. The last column shows the dimension η of the corresponding anisotropic polynomial space.

Theorem 3, namely

$$p_n = \frac{g_{min}}{g_n} p, \qquad p \in \mathbb{N}_+$$

where the decay coefficients g_n have been estimated according to results given in Lemma 2 and equation (9), namely:

$$g_n = \log(1 + r_n + \sqrt{r_n^2 + 2r_n}) \quad \text{and}$$
$$r_0 = \frac{a_{min}}{2\sqrt{3\lambda_0}}, \qquad r_n = r_{n_f+n} = \frac{a_{min}}{2\sqrt{3\lambda_n}}, \quad n = 1, \ldots, n_f.$$

Finally, a_{min} has been estimated as

$$a_{min} = c_0 - \sigma\sqrt{3\lambda_0} - \sum_{i=1}^{n_f} 2\sigma\sqrt{3\lambda_i}.$$

The convergence plot in Figure 7 (label TP) has been obtained taking $p = 1, 2, 3, 4$. The polynomial degrees used in the 7 directions, for different values of p, are reported in Table 1, together with the dimension η of the corresponding polynomial space.

Isotropic sparse grid Stochastic Collocation. The second methodology considered is the Stochastic Collocation method described in Section 6.2 based on isotropic sparse grids using Clenshaw-Curtis abscissas. Figure 6 shows the projection of the level 5 sparse grid onto the first 3 directions. The convergence plot in Figure 7 (label SC) has been obtained taking the sparse grids of level $w = 1, 2, 3, 4$. The number of collocation points in the four grids are $\eta = 22, 225, 1450, 6819$, respectively.

Monte Carlo Sampling. The third method considered is the classical Monte Carlo sampling. Here we have considered an increasing number of sample points

$$\eta = 100, 200, 400, 800, 1600, 3200,$$

and computed the sample average of the quantity of interest. The convergence curve is shown in Figure 7 (label MC). We have actually repeated the error

analysis for 20 independent replica. The solid line in Figure 7 corresponds to the average of the 20 errors (in absolute value) observed for each choice of η, while the dashed line corresponds the maximum error observed among the 20 replica.

Point Collocation. The fourth method considered is the so called *Point Collocation*, see [14], which is a non intrusive method as the Stochastic Collocation. It consists in randomly sampling the quantity of interest ψ in M points and seeking a polynomial approximation $\psi_p \in \mathcal{P}_p(\Gamma)$ of total degree p by a discrete least square approximation of the M sampled values (for details see [14]). The number of degrees of freedom of ψ_p is $\tilde{\eta} = \frac{(N+p)!}{N!p!}$ and we have chosen $M = 3\tilde{\eta}$. The total cost of each simulation (number of deterministic problems to solve) is therefore $\eta = 3\tilde{\eta}$. The convergence curve in Figure 7 (label PC) has been obtained taking $p = 1, 2, 3, 4, 5, 6$, which correspond to a total cost $\eta = 24, 108, 360, 990, 2376, 5148$, respectively. As for the Monte Carlo method, we have repeated the analysis for 20 independent replica and plotted the average error (solid line) and the maximum error (dashed line).

Monte Carlo Sampling + Point Collocation. The last method considered consists in taking the Point Collocation as a control variate of the Monte Carlo sample average to reduce its variance. In this case, we have used $M = 3\frac{(N+p)!}{N!p!}$ samples to generate the Point Collocation approximation and M (independent) samples in the Monte Carlo method with control variate. Therefore, the total cost of each simulation is $\eta = 2M = 6\frac{(N+p)!}{N!p!}$. In the convergence plot shown in Figure 7 (label MCPC) we have selected $p = 1, 2, 3, 4, 5, 6$ and shown the average error (solid line) and the maximum error (dashed line) obtained over 20 independent replica.

We see from this convergence analysis that, since the number of input random variables determining the diffusion coefficient is just $N = 7$, polynomial approximations obtained with either Stochastic Galerkin, Stochastic Collocation or Point Collocation seem to behave very similarly, clearly outperforming the Monte Carlo method. At the same time, the combination of Point Collocation and Monte Carlo reduces the original variance and yields a faster convergence than Monte Carlo. However, this procedure involves more deterministic solutions than the corresponding Point Collocation. The numerical results indicate that in this low input dimensional case, it is better to invest the extra work in the original Point Collocation method rather than on the combination of Point Collocation and Monte Carlo.

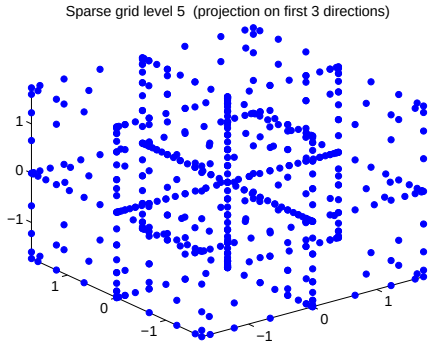

Figure 6: Projection of the level 5 isotropic sparse grid based on Clenshaw-Curtis knots onto the first three directions

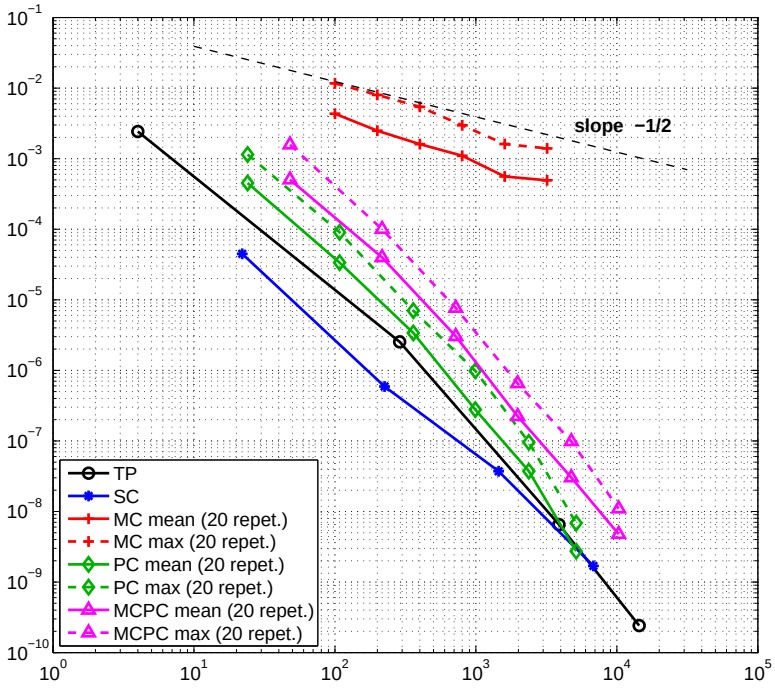

Figure 7: Convergence of different approximations with respect to η.

Acknowledgments.

The first and second authors were partially supported by the University of Austin Subcontract (Project Number 024550, Center for Predictive Computational Science). The second author also acknowledges his Dahlquist fellowship at the Royal Institute of Technology in Stockholm, Sweden and his start up funds at SC, Florida State University. He also wants to acknowledge the support of UdelaR in Uruguay.

References

[1] I. M. Babuška, R. Tempone, and G. E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42(2):800–825, 2004.

[2] I. M. Babuška, R. Tempone, and G. E. Zouraris. Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Comput. Methods Appl. Mech. Engrg.*, 194(12-16):1251–1294, 2005.

[3] I.M. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Numer. Anal.*, 45(3):1005–1034, 2007.

[4] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000.

[5] C. Canuto and T. Kozubek. A fictitious domain approach to the numerical solution of PDEs in stochastic domainsc. *Numer. Math.*, accepted for publication.

[6] C. W. Clenshaw and A. R. Curtis. A method for numerical integration on an automatic computer. *Numer. Math.*, 2:197–205, 1960.

[7] M.-K. Deb, I. M. Babuška, and J. T. Oden. Solution of stochastic partial differential equations using Galerkin finite element techniques. *Comput. Methods Appl. Mech. Engrg.*, 190:6359–6372, 2001.

[8] L. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. AMS, 1998.

[9] P. Frauenfelder, C. Schwab, and R. A. Todor. Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194(2-5):205–228, 2005.

[10] B. Ganapathysubramanian and N. Zabaras. Sparse grid collocation schemes for stochastic natural convection problems. *Journal of Computational Physics*, 225(1):652–685, 2007.

[11] R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach*. Springer-Verlag, New York, 1991.

[12] M. Grigoriu. *Stochastic calculus*. Birkhäuser Boston Inc., Boston, MA, 2002. Applications in science and engineering.

[13] W. Gui and I. Babuška. The h, p and h-p versions of the finite element method in 1 dimension. I. The error analysis of the p-version. *Numer. Math.*, 49(6):577–612, 1986.

[14] S. Hosder, R. Walters, and M. Balch. Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 2007.

[15] O. Knio and O. Le Maître. Uncertainty propagation in CFD using polynomial chaos decompositions. *Fluid Dynamics Research*, 38(9):616–640, 2006.

[16] P. Lévy. *Processus stochastiques et mouvement Brownien*. Éditions Jacques Gabay, 1992.

[17] O. P. Le Maître, O. M. Knio, H. N. Najm, and R. G. Ghanem. Uncertainty propagation using Wiener-Haar expansions. *J. Comput. Phys.*, 197(1):28–57, 2004.

[18] O. P. Le Maître, H. N. Najm, R. G. Ghanem, and O. M. Knio. Multi-resolution analysis of Wiener-type uncertainty propagation schemes. *J. Comput. Phys.*, 197(2):502–531, 2004.

[19] O. P. Le Maître, H. N. Najm, P. P. Pébay, R. G. Ghanem, and O. M. Knio. Multi-resolution-analysis scheme for uncertainty quantification in chemical systems. *SIAM J. Sci. Comput.*, 29(2):864–889 (electronic), 2007.

[20] L. Mathelin, M. Y. Hussaini, and T. A. Zang. Stochastic approaches to uncertainty quantification in CFD simulations. *Numer. Algorithms*, 38(1-3):209–236, 2005.

[21] H. G. Matthies and A. Keese. Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Comput. Methods Appl. Mech. Engrg.*, 194(12-16):1295–1331, 2005.

[22] F. Nobile, R. Tempone, and C. G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Num. Anal.*, 46(5):2411–2442, 2008.

[23] F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Num. Anal.*, 46(5):2309–2345, 2008.

[24] B. Øksendal. *Stochastic Differential Equations. An introduction with applications*. Springer–Verlag, fifth edition, 1998.

[25] F. Riesz and B. Sz.-Nagy. *Functional analysis*. Dover, 1990.

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

[27] M.A. Tatang. *Direct incorporation of uncertainty in chemical and environmental engineering systems*. PhD thesis, MIT, 1995.

[28] R. A. Todor and C. Schwab. Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA J Numer Anal*, 27(2):232–261, 2007.

[29] X. Wan and G. E. Karniadakis. Beyond Wiener-Askey expansions: handling arbitrary PDFs. *J. Sci. Comput.*, 27(1-3):455–464, 2006.

[30] X. Wan and G. E. Karniadakis. Multi-element generalized polynomial chaos for arbitrary probability measures. *SIAM J. Sci. Comput.*, 28(3):901–928, 2006.

[31] G. W. Wasilkowski and H. Woźniakowski. Explicit cost bounds of algorithms for multivariate tensor product problems. *Journal of Complexity*, 11:1–56, 1995.

[32] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.

[33] D. Xiu and G. E. Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Comput. Methods Appl. Mech. Engrg.*, 191(43):4927–4948, 2002.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

22/2008 F. Nobile, R. Tempone:
Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients

21/2008 P. Antonietti, E. Süli:
Domain Decomposition Preconditioning for Discontinuous Galerkin Approximations of Convection-Diffusion Problems

20/2008 F. David, S. Micheletti, S. Perotto:
Model adaption enriched with an anisotropic mesh spacing for advection-diffusion-rection systems

19/2008 S. Badia, F. Nobile, C. Vergara:
Robin-Robin preconditioned Krylov methods for fluid-structure interaction problems

18/2008 L. Bonaventura, S. Castruccio, P. Crippa, G. Lonati:
Geostatistical estimate of PM10 concentrations in Northern Italy: validation of kriging reconstructions with classical and flexible variogram models

17/2008 A. Ern, S. Perotto, A. Veneziani:
Hierarchical model reduction for advection-diffusion-reaction problems

16/2008 L. Formaggia, E. Miglio, A. Mola, A. Scotti:
Numerical simulation of the dynamics of boat by a variational inequality approach

15/2008 S. Micheletti, S. Perotto:
An anisotropic mesh adaptation procedure for an optimal control problem of the advection-diffusion-reaction equation

14/2008 C. D'Angelo, P. Zunino:
A finite element method based on weighted interior penalties for heterogeneous incompressible flows

13/2008 L.M. Sangalli, P. Secchi, S. Vantini, V. Vitelli:
K-means alignment for curve clustering