ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

MATHICSE

Mathematics Institute of Computational Science and Engineering

School of Basic Sciences - Section of Mathematics

MATHICSE Technical Report

Nr. 40.2014

September 2014

A posteriori error estimations for elliptic partial differential equations with small uncertainties

Diane Guignard, Fabio Nobile, Marco Picasso

http://mathicse.epfl.ch

Phone: +41 21 69 37648

Address:
EPFL - SB - MATHICSE (Bâtiment MA)
Station 8 - CH-1015 - Lausanne - Switzerland

Fax: +41 21 69 32545

A posteriori error estimations for elliptic partial differential equations with small uncertainties.

Diane Guignard[1], Fabio Nobile[1], and Marco Picasso[1]

[1] MATHICSE, Ecole Polytechnique Fédérale Lausanne, Station 8, CH 1015, Lausanne, Switzerland
Correspondence to: diane.guignard@epfl.ch.

July 21, 2014

Abstract

In this paper, a finite element error analysis is performed on a class of linear and nonlinear elliptic problems with small uncertain input. Using a perturbation approach, the exact (random) solution is expanded up to a certain order with respect to a parameter that controls the amount of randomness in the input and discretized by finite elements. We start by studying a diffusion (linear) model problem with a random coefficient characterized via a finite number of random variables. *A priori* and *a posteriori* estimates of the error between the exact and approximate solution are given in various norms, including goal-oriented error estimation. The analysis is then extended to a class of nonlinear problems. We finally illustrate the theoretical results through numerical examples, along with a comparison with the Stochastic Collocation method in terms of computational costs.

1 Introduction

Partial differential equations (PDEs) are widely used for modelling problems in many fields such as physics, biology or engineering. Nowadays, uncertainty is often included in mathematical models arising from the simulation of complex systems. The uncertainty can reflect an intrinsic variability of the system or our inability to adequately characterize all the input, due for instance to experimental measurements. It can occur in the input data, the geometry, the boundary conditions, the initial condition or combinations of them. One way to model such uncertainties is to use probability theory, characterizing the uncertainties by random variables or more generally by random fields.

Much effort has thus been put into the development of methods for solving PDEs with random input. Other than Monte-Carlo (MC) type methods, which regroup among others the standard MC (see [1] for instance), the quasi-MC [2, 3] and the multi-level MC [4–6] methods, we can mention the stochastic spectral methods comprising the Stochastic Galerkin (SG) [7–11] and the Stochastic Collocation (SC) [9, 12–15] methods. These methods exploit the possible regularity of the solution with respect to the random input combining the generalized Polynomial Chaos (gPC) expansion of the solution with a Galerkin projection or an interpolation procedure. In both methods, an approximation in the physical space can be obtained using for instance the Finite Element (FE) method. An *a posteriori* error estimate in the *energy* norm for the SG-FEM is derived in [16, 17], where adaptive refinement algorithms are proposed for both stochastic and physical spaces. In the algorithm proposed in [17], the refined mesh is the same for all generalized polynomial chaos (gPC) modes, contrary to the one in [16] where the refinement procedure is applied independently for each mode. For the SC-FEM method, an *a priori* error estimation is given in [12] but, to our knowledge, no *a posteriori* error estimator for the whole solution in suitable norms has been derived yet. Recently, *a posteriori* error estimations for a specific quantity of interest have been developed.

Goal-oriented error estimators can be found in [18–20] for the SG method and in [21] for the SC method.

This work is focused on PDEs with small uncertainties (for instance the linear model problem $-\operatorname{div}(a\nabla u) = f$ with $a = a_0 + \varepsilon(a_1 Y_1 + ... + a_L Y_L)$ where ε is small and $Y_1, ..., Y_L$ are random variables). We therefore follow a different path and adopt a perturbation approach (see e.g. [22,23]) expanding the stochastic solution u as

$$u(\mathbf{x}, \omega) := u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \omega) + \mathcal{O}(\varepsilon^2) \tag{1}$$

where ε is a parameter controlling the magnitude of uncertainty in the input which is assumed to be small. Uncoupled problems can be derived to find the deterministic part u_0 and the stochastic one u_1 (and higher order terms), the error analysis being performed in various norms. The main goal of this paper is then to derive *a posteriori* estimates for the error between the exact (random) solution u and certain approximations to be defined. For instance, if we write $u_{0,h}$ the FE approximation of u_0, then we will show that the error $u - u_{0,h}$ splits into two parts. More precisely, we will derive an *a posteriori* error estimator of the form

$$\|u - u_{0,h}\| \leq C \left(\eta_1(h)^2 + \eta_2(\varepsilon)^2\right)^{\frac{1}{2}},$$

with the norm $\|\cdot\|$ to be defined and where η_1 and η_2 are deterministic quantities that depend only on $u_{0,h}$ and the input data. Therefore, by solving only one deterministic problem we directly know how much of the error is due to the space discretization and how much is due to the uncertainty. This information is given respectively by η_1 and η_2. This estimator, easy and cheap to compute, can then be used to determine a mesh size yielding comparable accuracy in h and ε. The next term $u_{1,h}$ (and then higher order terms) can then be added to get better accuracy in ε.

We mention that the *a posteriori* error estimator that we obtain for $u - u_{0,h}$ in this work have similarities with the one derived in [24], although the context of the two papers is quite different. In [24] the authors derive an adaptive finite element method for elliptic PDEs with discontinuous coefficients. The proposed algorithm takes into account the error due to FE approximation but also the effect of replacing the discontinuous input data by some piecewise polynomial approximation, which plays the same role as a_0 in our setting. More precisely, before applying a standard AFEM to the problem, the mesh is first refined so that the discontinuous input are approximated by piecewise polynomials with a prescribed accuracy. The specific form of the uncertain input we consider here, see (3), allows us to increase the accuracy in ε by adding terms in the expansion (1) of u.

This paper is organized as follows. In Section 2, a diffusion model problem with homogeneous Dirichlet boundary conditions and random diffusion coefficient is studied. The diffusion coefficient is assumed, among others, to be expanded as a finite sum which depends on independent random variables of zero mean and finite variance. Error analysis in H_0^1 and L^2 norms in the physical domain, as well as goal-oriented error estimation, is performed when the exact (random) solution u is approximated by the (deterministic) FE approximation of u_0. Then, the error between u and the FE approximation of $u_0 + \varepsilon u_1$ is considered, before giving a generalization for an approximation of arbitrary order in ε. The theory is then extended to nonlinear problems in Section 3. In Section 4, a comparison of the computational costs for the Stochastic Collocation method and the one presented here is performed. Finally, Section 5 is devoted to numerical examples used to illustrate and validate the theoretical results.

2 A linear model problem

We first study a diffusion problem with random diffusion coefficient. Well-posedness and (*a priori*, *a posteriori*) error estimates are proved in several norms.

2.1 Problem setting

Let D be a bounded polyhedral domain in $\mathbb{R}^d$, $d = 1, 2, 3$, and $(\Omega, \mathcal{F}, P)$ a complete probability space, where 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. The following problem is considered

find $u : D \times \Omega \to \mathbb{R}$ such that $P-$almost everywhere (in other words almost surely):

$$\left\{ \begin{array}{rcll} -\operatorname{div}\left(a(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)\right) & = & f(\mathbf{x}) & \text{in } D \\ u(\mathbf{x},\omega) & = & 0 & \text{on } \partial D. \end{array} \right. \tag{2}$$

For simplicity, the right-hand side f is supposed to be deterministic, $f \in L^2(D)$, but the case of stochastic forcing term could be considered as well adding no real difficulty. Note that the divergence and gradient operators apply only on $\mathbf{x}$, the physical space variable.

The following assumptions (see [7, 12, 25] for instance) are made on the random diffusion coefficient a to ensure, among others, the well-posedness of the problem.

(A1) coercivity and continuity: a is bounded and uniformly coercive, i.e. there exist two real constants $0 < a_{min} \leq a_{max} < \infty$ such that

$$P(a_{min} \leq a(\mathbf{x},\omega) \leq a_{max}, \forall \mathbf{x} \in \overline{D}) = 1.$$

(A2) finite dimensional noise: a can be parametrized with L mutually independent random variables $a(\mathbf{x},\omega) = a(\mathbf{x}, Y_1(\omega), Y_2(\omega), ..., Y_L(\omega))$. More precisely, we assume that a can be expanded as

$$a(\mathbf{x},\omega) = a_0(\mathbf{x}) + \varepsilon \sum_{j=1}^{L} a_j(\mathbf{x}) Y_j(\omega), \tag{3}$$

where the $(Y_j)_{j=1}^L$ are independent random variables of zero mean and finite variance σ^2. Assuming $a_j \in L^\infty(D)$ for $j = 0, 1, ..., L$ is enough to ensure the well-posedness of the problem; however, in what follows, we will assume more regularity, namely $a_j \in W^{1,\infty}(D)$, $j = 0, ..., L$, in order to avoid difficulties that are beyond the scope of this paper. We refer to [26] for a derivation of *a posteriori* error estimation in the case of discontinuous coefficient.

Notice that as a consequence of assumption (A1), the random variables Y_j, $j = 1, ..., L$, have to be bounded almost surely.

Remark 2.1. *The characterization (3) of the random input can be achieved using, for instance, a (truncated) Karhunen-Loève type expansion [27, 28]. Note that in this paper, we do not take into account the error made when the random input is approximated via a finite number of random variables, i.e. we assume that the random diffusion coefficient of problem (2) is accurately described by (3).*

The stochasticity of the problem can therefore be parametrized by the random vector $\mathbf{y}(\omega) = (Y_1(\omega), ..., Y_L(\omega))$. For $j = 1, ..., L$, let Γ_j denote the bounded image in $\mathbb{R}$ of the random variable Y_j, i.e. $\Gamma_j := Y_j(\Omega)$, and write $\rho_j : \Gamma_j \to \mathbb{R}^+$ its probability density function. Thanks to the independence of the random variables, the joint density function $\rho : \Gamma \to \mathbb{R}^+$ of the random vector $\mathbf{y}$ is then given by $\rho(\mathbf{y}) = \prod_{j=1}^L \rho_j(Y_j)$, where $\Gamma = \Gamma_1 \times \Gamma_2 \times \cdots \times \Gamma_L$. By definition, the expected value of any measurable function $g : \Gamma \to \mathbb{R}$ is then $\mathbb{E}[g] = \int_\Gamma g(\mathbf{y})\rho(\mathbf{y})d\mathbf{y}$. The probability space $(\Omega, \mathcal{F}, P)$ can thus be replaced by $(\Gamma, B(\Gamma), \rho(\mathbf{y})d\mathbf{y})$, where $B(\Gamma)$ is the Borel σ-algebra on Γ and $\rho(\mathbf{y})d\mathbf{y}$ is the distribution measure of the random vector $\mathbf{y}$ [29]. Due to the Doob-Dynkin lemma (see [25, p.6] for instance), the solution u is a function of the random variables Y_j, i.e. $u(\mathbf{x},\omega) = u(\mathbf{x}, \mathbf{y}(\omega))$.

The stochastic elliptic boundary value problem (2) can now be written in the following deterministic parametric form

find $u : D \times \Gamma \to \mathbb{R}$ such that for almost every $\mathbf{y} \in \Gamma$ we have

$$\left\{ \begin{array}{rcll} -\operatorname{div}\left(a(\mathbf{x},\mathbf{y})\nabla u(\mathbf{x},\mathbf{y})\right) & = & f(\mathbf{x}), & \mathbf{x} \in D \\ u(\mathbf{x},\mathbf{y}) & = & 0, & \mathbf{x} \in \partial D. \end{array} \right. \tag{4}$$

The weak solution of problem (4) can either be seen as a function in the tensor space $H_0^1(D) \otimes L_\rho^2(\Gamma)$ or in the Bochner space $L_\rho^2(\Gamma; H_0^1(D))$, which are isomorphic (see [7]). We adopt here the second point of view. Let $H_0^1(D)$ be endowed with the seminorm $H^1(D)$, i.e.

$$\|v\|_{H_0^1(D)} := \|\nabla v\|_{L^2(D)} = \left(\int_D |\nabla v|^2 \right)^{\frac{1}{2}}.$$

The point-wise weak form of problem (4) reads

find $u(\cdot, \mathbf{y}) \in H_0^1(D)$ such that

$$\mathcal{A}(u(\cdot, \mathbf{y}), v; \mathbf{y}) = \mathcal{F}(v) \quad \forall v \in H_0^1(D), \text{ for almost every } \mathbf{y} \in \Gamma, \tag{5}$$

where

$$\mathcal{A}(u(\cdot, \mathbf{y}), v; \mathbf{y}) = \int_D a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}) d\mathbf{x}, \tag{6}$$

$$\mathcal{F}(v) = \int_D f(\mathbf{x}) v(\mathbf{x}) d\mathbf{x}. \tag{7}$$

Thanks to Lax-Milgram's lemma, the coercivity and continuity assumptions on a ensure the well-posedness of problem (5). Indeed, since a is bounded from below and above almost surely, the bilinear form $\mathcal{A}$ is continuous and coercive with constant of continuity and coercivity given respectively by a_{max} and a_{min}. Furthermore, the linear (deterministic) functional $\mathcal{F}$ is continuous, with constant of continuity equal to $C_P \|f\|_{L^2(D)}$, where C_P denotes the constant in the Poincaré inequality. Therefore, the random solution u of problem (5) satisfies almost surely

$$\|\nabla u\|_{L^2(D)} \leq \frac{C_P}{a_{min}} \|f\|_{L^2(D)}. \tag{8}$$

From now on, we assume ε in (3) to be small and expand the solution u with respect to ε up to a certain order $N \in \mathbb{N}$

$$u(\mathbf{x}, \mathbf{y}(\omega)) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{y}(\omega)) + ... + \varepsilon^N u_N(\mathbf{x}, \mathbf{y}(\omega)) + \mathcal{O}(\varepsilon^{N+1}). \tag{9}$$

Inserting the latter expansion into (4) and keeping the $\mathcal{O}(1)$ term with respect to ε yields the problem

find $u_0 : D \to \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_0(\mathbf{x}) \nabla u_0(\mathbf{x})) &= f(\mathbf{x}), \quad \mathbf{x} \in D \\ u_0(\mathbf{x}) &= 0, \quad \mathbf{x} \in \partial D. \end{cases} \tag{10}$$

Then, searching $u_1(\mathbf{x}, \mathbf{y}(\omega)) = \sum_{j=1}^L U_j(\mathbf{x}) Y_j(\omega)$ and keeping $\mathcal{O}(\varepsilon)$ terms in (4) yields the L problems

find $U_j : D \to \mathbb{R}$ such that

$$\begin{cases} -\operatorname{div}(a_j(\mathbf{x}) \nabla u_0(\mathbf{x}) + a_0(\mathbf{x}) \nabla U_j(\mathbf{x})) &= 0, \quad \mathbf{x} \in D \\ U_j(\mathbf{x}) &= 0, \quad \mathbf{x} \in \partial D \end{cases} \quad j = 1, ..., L, \tag{11}$$

in which the solution u_0 of problem (10) is needed. Notice that for $j = 1, ..., L$, the function U_j is related to $\frac{\partial u(\mathbf{x}, \mathbf{y}_0)}{\partial y_j}$ with $\mathbf{y}_0 = \mathbf{0}$. Similarly, we can use the solutions U_j, $j = 1, ..., L$, of problem (11) to compute the deterministic part of the next term in the expansion (9), which in

turn is related to the second derivatives $\frac{\partial^2 u(\mathbf{x},\mathbf{y}_0)}{\partial y_j \partial y_k}$, $j,k = 1,...,L$. Indeed, if we write $u_2(\mathbf{x},\mathbf{y}(\omega)) = \sum_{j,k=1}^{L} U_{jk}(\mathbf{x}) Y_j(\omega) Y_k(\omega)$, keeping the $\mathcal{O}(\varepsilon^2)$ terms in (4), we get the L^2 problems

find $U_{jk} : D \to \mathbb{R}$ such that

$$\left\{ \begin{array}{rcll} -\operatorname{div}\left(a_j(\mathbf{x})\nabla U_k(\mathbf{x}) + a_0(\mathbf{x})\nabla U_{jk}(\mathbf{x})\right) & = & 0, & \mathbf{x} \in D \\ U_{jk}(\mathbf{x}) & = & 0, & \mathbf{x} \in \partial D \end{array} \right. \quad j,k = 1,...,L. \tag{12}$$

Remark 2.2. *We will prove in sections 2.2, 2.3 and 2.4 that*

$$u - u_0 = \mathcal{O}(\varepsilon), \quad u - (u_0 + \varepsilon u_1) = \mathcal{O}(\varepsilon^2) \quad \textit{and} \quad u - (u_0 + \varepsilon u_1 + \varepsilon^2 u_2) = \mathcal{O}(\varepsilon^3).$$

The solution to the deterministic problems (10), (11) and (12) can be approximated using for instance the finite element method. For any $h > 0$, let $\mathcal{T}_h$ be a family of partitions of D into d-simplices (intervals, triangles, tetrahedra) K of diameter $h_K \leq h$. Unless otherwise stated, we will always consider shape regular (see [30]) meshes of D, i.e. decompositions such that there exists a constant $c > 0$ satisfying

$$\frac{h_K}{\rho_K} \leq c \quad \forall K \in \mathcal{T}_h, \forall h > 0$$

where $\rho_K = \sup\{diam(B) : B \text{ is a ball contained in } K\}$. Let $V_h \subset H_0^1(D)$ be the space of continuous, piecewise linear finite element functions associated to $\mathcal{T}_h$ that vanish on ∂D, that is

$$V_h := \{v_h \in C^0(\bar{D}) : v_{h\,|K} \in \mathbb{P}_1 \ \forall K \in \mathcal{T}_h\} \cap H_0^1(D),$$

where $\mathbb{P}_1$ is the set of polynomials of degree less or equal to 1.

In the derivation of *a priori* and *a posteriori* error estimates, we will need an interpolant operator which maps $H_0^1(D)$ to V_h, along with interpolation error bounds. We distinguish the cases $d = 1$ and $d = 2, 3$. For the one-dimensional case, any function of $H_0^1(D)$ is continuous thanks to Sobolev embedding theorem. Therefore, the Lagrange interpolant operator $r_h : C^0(\bar{D}) \to V_h$, which requires point value evaluations, is well-defined and satisfies the following error bounds: there exists a constant $C > 0$ such that $\forall h > 0$, $\forall K \in \mathcal{T}_h$ and all $v \in H_0^1(D)$ we have

$$\|v - r_h v\|_{L^2(K)} \leq C h_K \|v'\|_{L^2(K)} \tag{13}$$

and for all $v \in H^2(D)$

$$\|v - r_h v\|_{L^2(K)} + h_K \|v' - (r_h v)'\|_{L^2(K)} \leq h_K^2 \|v''\|_{L^2(K)}.$$

For the case $d = 2, 3$, the functions of $H^2(D)$ are continuous and we have the following error bound (see [30, 31] for instance) based on the Bramble-Hilbert lemma: there exists a constant $C > 0$ such that $\forall h > 0$, $\forall K \in \mathcal{T}_h$ and all $v \in H^2(K)$ we have

$$\|v - r_h v\|_{L^2(K)} + h_K \|\nabla(v - r_h v)\|_{L^2(K)} \leq C h_K^2 |v|_{H^2(K)}. \tag{14}$$

In general however, such regularity might not be reached by the solution of problem (5), since we are seeking for a solution in $H_0^1(D)$ in the physical space. In that case, we will use the Clément interpolant [32] operator $\mathcal{I}_h : H_0^1(D) \to V_h$ which satisfies the following interpolation results

$$\|v - \mathcal{I}_h v\|_{L^2(K)} \leq C h_K |v|_{H^1(N(K))} \tag{15}$$

and

$$\|v - \mathcal{I}_h v\|_{L^2(e)} \leq C h_e^{\frac{1}{2}} |v|_{H^1(N(K_e))}, \tag{16}$$

where, for an internal edge e, K_e is the union of the two elements touching e and $N(K)$ (respectively $N(K_e)$) denotes the patch of elements associated to K (respectively K_e).

We will now derive *a priori* and *a posteriori* error estimators in various norms, the error being the difference between the exact solution and a certain approximate solution to be defined. We first start by giving error estimates between the exact solution u and $u_{0,h}$, the FE approximation of u_0. Our goal is to decompose the error into two parts, the error due to the finite element approximation (h) and the error due to the uncertainty (ε).

2.2 First order approximation

We consider the case where $N = 0$ in the expansion (9). In that case, the error due to the stochastic truncation is of order ε. Indeed, for any $v \in H_0^1(D)$ we have almost surely

$$\int_D a\nabla(u - u_0) \cdot \nabla v = \int_D fv - \int_D a\nabla u_0 \cdot \nabla v = -\varepsilon \sum_{j=1}^{L} Y_j \int_D a_j \nabla u_0 \cdot \nabla v. \tag{17}$$

Using the FEM, the unknown solution u_0 of problem (10) is approximated by $u_{0,h}$, the solution of

$$\text{find } u_{0,h} \in V_h \text{ such that } \int_D a_0 \nabla u_{0,h} \cdot \nabla v_h = \int_D f v_h \quad \forall v_h \in V_h. \tag{18}$$

The next section is devoted to *a priori* error estimation for the strong and weak error, which give information on the asymptotic behaviour of the error. In particular, we will show that the order of the error of the mean in ε is twice the order of the strong error, while the order of the error in h is the same for both. Sections 2.2.2, 2.2.3 and 2.2.4 are instead devoted to *a posteriori* error estimates in different norms.

2.2.1 A priori error estimate

Strong error estimate

Let us first give an error estimate on the strong error, i.e. on the error between u and $u_{0,h}$ in the $L^2_\rho(\Gamma; H_0^1(D))$ norm. Under suitable regularity assumptions on u_0 and constraint on the a_j, we have the following *a priori* error estimator.

Proposition 2.3. *Let u, u_0 and $u_{0,h}$ be the solution of problems (2), (10) and (18) respectively. Assume that for a fixed value $\alpha > \frac{1}{2}$, there exists a constant D_α such that $\sum_{j=1}^{L} \|a_j^2\|_{L^\infty(D)} j^{2\alpha} \leq D_\alpha$. If $u_0 \in H^2(D)$, then we have the a priori error estimate*

$$\mathbb{E}\left[\|\nabla(u - u_{0,h})\|^2_{L^2(D)}\right]^{\frac{1}{2}} \leq \sqrt{2}\left[\frac{a_{0,max}}{a_{0,min}} C^2 h^2 |u_0|^2_{H^2(D)} + C_\alpha D_\alpha \frac{\varepsilon^2 \sigma^2 C_P^2}{a_{0,min}^2 a_{min}^2} \|f\|^2_{L^2(D)}\right]^{\frac{1}{2}} \tag{19}$$

where $C > 0$ is the constant, independent of u, u_0, h and ε, that appear in (14) and C_α depends only on α. Therefore, there exists a constant $\tilde{C} > 0$ independent of h and ε such that

$$\mathbb{E}\left[\|\nabla(u - u_{0,h})\|^2_{L^2(D)}\right]^{\frac{1}{2}} \leq \tilde{C}(h + \varepsilon).$$

Proof. Using the fact that almost surely it holds

$$\int_D a_0 \nabla u_0 \cdot \nabla v = \int_D fv = \int_D a\nabla u \cdot \nabla v \quad \forall v \in V,$$

we have for any $v \in V$

$$\begin{aligned}
\int_D a_0 \nabla(u - u_{0,h}) \cdot \nabla v &= \int_D a_0 \nabla(u - u_0) \cdot \nabla v + \int_D a_0 \nabla(u_0 - u_{0,h}) \cdot \nabla v \\
&= -\int_D (a - a_0)\nabla u \cdot \nabla v + \int_D a_0 \nabla(u_0 - u_{0,h}) \cdot \nabla v \\
&\leq \left[\left(\int_D \frac{(a_0 - a)^2}{a_0} |\nabla u|^2\right)^{\frac{1}{2}} + \left(\int_D a_0 |\nabla(u_0 - u_{0,h})|^2\right)^{\frac{1}{2}}\right] \cdot \left(\int_D a_0 |\nabla v|^2\right)^{\frac{1}{2}}.
\end{aligned} \tag{20}$$

Thanks to the inequality $(a+b)^2 \leq 2(a^2+b^2)$, taking $v = u - u_{0,h} \in V$ for a.e. $\mathbf{y} \in \Gamma$ in the last inequality yields

$$\left(\int_D a_0 |\nabla(u-u_{0,h})|^2\right)^{\frac{1}{2}} \leq \sqrt{2}\left[\frac{1}{a_{0,min}}\int_D (a-a_0)^2|\nabla u|^2 + \int_D a_0|\nabla(u_0-u_{0,h})|^2\right]^{\frac{1}{2}}. \quad (21)$$

The second term of the right-hand side of (21) can be bounded in a standard manner as follows. Using the Galerkin orthogonality property

$$\int_D a_0 \nabla(u_0 - u_{0,h}) \cdot \nabla v_h = 0 \quad \forall v_h \in V_h,$$

we easily get

$$\int_D a_0 |\nabla(u_0-u_{0,h})|^2 \leq a_{0,max} \|\nabla(u_0 - \mathcal{I}_h u_0)\|^2_{L^2(D)}.$$

Since $u_0 \in H^2(D)$ by hypothesis, thanks to the interpolation result (14) we get

$$\int_D a_0 |\nabla(u_0-u_{0,h})|^2 \leq a_{0,max} C^2 h^2 |u_0|^2_{H^2(D)}. \quad (22)$$

Therefore, using this last relation and the lower bound for a_0 in (21) yields

$$\|\nabla(u-u_{0,h})\|^2_{L^2(D)} \leq 2\left[\frac{1}{a^2_{0,min}}\int_D (a-a_0)^2|\nabla u|^2 + \frac{a_{0,max}}{a_{0,min}} C^2 h^2 |u_0|^2_{H^2(D)}\right].$$

Then, we take the expected value on both side of the last inequality to get

$$\mathbb{E}\left[\|\nabla(u-u_{0,h})\|^2_{L^2(D)}\right] \leq 2\left[\frac{1}{a^2_{0,min}}\mathbb{E}\left[\int_D (a-a_0)^2|\nabla u|^2\right] + \frac{a_{0,max}}{a_{0,min}} C^2 h^2 |u_0|^2_{H^2(D)}\right]. \quad (23)$$

To complete the proof, we finally bound the expected value that appear on the right-hand side of (23). Using the relation $\sum_i a_i b_i \leq \left(\sum_i a_i^2\right)^{\frac{1}{2}} \left(\sum_i b_i^2\right)^{\frac{1}{2}}$, we have

$$(a-a_0)^2 = \varepsilon^2 \left(\sum_{j=1}^L a_j j^\alpha j^{-\alpha} Y_j\right)^2 \leq \varepsilon^2 \left(\sum_{j=1}^L a_j^2 j^{2\alpha}\right)\left(\sum_{j=1}^L Y_j^2 j^{-2\alpha}\right) \leq D_\alpha \varepsilon^2 \sum_{j=1}^L Y_j^2 j^{-2\alpha}.$$

Therefore, thanks to (8) and the fact that $\mathbb{E}[Y_j^2] = \sigma^2$, we obtain

$$\mathbb{E}\left[\int_D (a-a_0)^2|\nabla u|^2\right] \leq D_\alpha \frac{\varepsilon^2\sigma^2 C_P^2}{a^2_{min}} \|f\|^2_{L^2(D)} \sum_{j=1}^L j^{-2\alpha}.$$

Since $\alpha > \frac{1}{2}$, the series $\sum_{j=1}^\infty j^{-2\alpha}$ converges which concludes the proof. □

Remark 2.4. *Using the relation $(\sum_{j=1}^L x_j)^2 \leq L \sum_{j=1}^L x_j^2$, the expected value that appears in the right-hand side of (23) could also be bounded by*

$$\mathbb{E}\left[\int_D (a-a_0)^2|\nabla u|^2\right] \leq L \frac{\varepsilon^2\sigma^2 C_P^2}{a^2_{min}} \|f\|^2_{L^2(D)} \sum_{j=1}^L \|a_j^2\|_{L^\infty(D)}.$$

Although we do not need additional constrain on the functions a_j, this bound explodes when L tends to infinity. Moreover, we notice that starting the analysis with a instead of a_0 in (20) alleviates the troubles encountered to bound (23), but this is not the natural way to perform a priori error analysis.

Mean of the error estimate

We are now interested in the error on the law of u. We restrict, in particular, to the $H_0^1(D)$-norm of the expected value of $u - u_{0,h}$. In this case, the statistical error is of order 2, to be compared to the order 1 of the strong error. Under the same regularity condition on u_0, we can show the following *a priori* error estimate.

Proposition 2.5. *Let u, u_0 and $u_{0,h}$ be the solution of problems (2), (10) and (18) respectively. If $u_0 \in H^2(D)$, then we have the a priori error estimate*

$$\|\mathbb{E}\left[u - u_{0,h}\right]\|_{H_0^1(D)} \leq \sqrt{\frac{a_{0,max}}{a_{0,min}}} C_1 h |u_0|_{H^2(D)} + \frac{\varepsilon^2 \sigma^2 C_P}{a_{0,min}^3} \|f\|_{L^2(D)} \sum_{j=1}^{L} \|a_j\|_{L^\infty(D)}^2 + C_2 \varepsilon^3, \quad (24)$$

where $C_1 > 0$ is the constant in (14) and C_2 is a constant independent of u, h and ε. Therefore, there exists a constant $\tilde{C} > 0$ independent of h and ε such that

$$\|\mathbb{E}\left[u - u_{0,h}\right]\|_{H_0^1(D)} \leq \tilde{C}(h + \varepsilon^2).$$

Proof. Let us define $u_1 = \sum_{j=1}^{L} U_j Y_j$, where U_j is the solution of problem (11) for $j = 1, ..., L$. First, the expected value of the error $u - u_{0,h}$ naturally splits into two parts

$$\mathbb{E}[u - u_{0,h}] = \mathbb{E}[u - u_0] + (u_0 - u_{0,h})$$

and thus, thanks to the triangle inequality, we get

$$\|\mathbb{E}\left[u - u_{0,h}\right]\|_{H_0^1(D)} \leq \|\mathbb{E}[u - u_0]\|_{H_0^1(D)} + \|u_0 - u_{0,h}\|_{H_0^1(D)}.$$

From (22), we deduce a bound for the second term given by

$$\|u_0 - u_{0,h}\|_{H_0^1(D)} \leq \sqrt{\frac{a_{0,max}}{a_{0,min}}} C_1 h |u_0|_{H^2(D)},$$

where C_1 is the constant that appears in (14). Let us bound the term $\|\mathbb{E}[u - u_0]\|_{H_0^1(D)}$, which is due to the uncertainty in the diffusion coefficient. Proceeding as in (17), we can easily show (see (42) for more details) that for any $v \in V$ we have a.s.

$$\int_D a \nabla(u - (u_0 + \varepsilon u_1)) \cdot \nabla v = -\varepsilon^2 \sum_{i,j=1}^{L} Y_i Y_j \int_D a_i \nabla U_j \cdot \nabla v \quad (25)$$

using the fact that $\int_D (a_j \nabla u_0 + a_0 \nabla U_j) \cdot \nabla v = 0$ for all $v \in V$. Therefore, we have

$$\int_D a_0 \nabla(u - (u_0 + \varepsilon u_1)) \cdot \nabla v = -\int_D (a - a_0) \nabla(u - (u_0 + \varepsilon u_1)) \cdot \nabla v - \varepsilon^2 \sum_{i,j=1}^{L} Y_i Y_j \int_D a_i \nabla U_j \cdot \nabla v.$$

Since $\mathbb{E}[u_1] = 0$ and $\mathbb{E}[Y_i Y_j] = \sigma^2 \delta_{ij}$, taking the expected value on both sides of last equality yields

$$\int_D a_0 \nabla \mathbb{E}[u - u_0] \cdot \nabla v = \mathbb{E}\left[-\int_D (a - a_0) \nabla(u - (u_0 + \varepsilon u_1)) \cdot \nabla v\right] - \varepsilon^2 \sigma^2 \sum_{j=1}^{L} \int_D a_j \nabla U_j \cdot \nabla v.$$

Thanks to Jensen's inequality, we obtain

$$\int_D a_0 \nabla \mathbb{E}[u - u_0] \cdot \nabla v \leq \mathbb{E}\left[\|a - a_0\|_{L^\infty(D)} \|\nabla(u - (u_0 + \varepsilon u_1))\|_{L^2(D)}\right] \|\nabla v\|_{L^2(D)}$$
$$+ \varepsilon^2 \sigma^2 \|\nabla v\|_{L^2(D)} \sum_{j=1}^{L} \|a_j\|_{L^\infty(D)} \|\nabla U_j\|_{L^2(D)}.$$

If we take $v = \mathbb{E}[u - u_0]$ in the last inequality, we get

$$\|\mathbb{E}[u-u_0]\|_{H_0^1(D)} \leq \frac{1}{a_{0,min}}\{\mathbb{E}\left[\|a-a_0\|_{L^\infty(D)}\|\nabla(u-(u_0+\varepsilon u_1))\|_{L^2(D)}\right] + \varepsilon^2\sigma^2\sum_{j=1}^{L}\|a_j\|_{L^\infty(D)}\|\nabla U_j\|_{L^2(D)}\}. \tag{26}$$

We now give a bound on $\|\nabla U_j\|_{L^2(D)}$, $j = 1, ..., L$. First, using standard techniques (Cauchy-Schwarz, Poincaré inequalities, lower bound for a_0), we get the following bound on the solution of problem (10)

$$\|\nabla u_0\|_{L^2(D)} \leq \frac{C_P}{a_{0,min}}\|f\|_{L^2(D)}.$$

Then, taking $v = U_j$ as test function in the weak formulation of problem (11) yields

$$a_{0,min}\|\nabla U_j\|^2_{L^2(D)} \leq \int_D a_0|\nabla U_j|^2 = -\int_D a_j\nabla u_0\cdot\nabla U_j \leq \|a_j\|_{L^\infty(D)}\|\nabla u_0\|_{L^2(D)}\|\nabla U_j\|_{L^2(D)}$$

and thus

$$\|\nabla U_j\|_{L^2(D)} \leq \frac{C_P}{a^2_{0,min}}\|f\|_{L^2(D)}\|a_j\|_{L^\infty(D)}.$$

Inserting this result in (26), we get

$$\|\mathbb{E}[u-u_0]\|_{H_0^1(D)} \leq \frac{1}{a_{0,min}}\{\mathbb{E}\left[\|a-a_0\|_{L^\infty(D)}\|\nabla(u-(u_0+\varepsilon u_1))\|_{L^2(D)}\right] + \frac{\varepsilon^2\sigma^2 C_P}{a^2_{0,min}}\|f\|_{L^2(D)}\sum_{j=1}^{L}\|a_j\|^2_{L^\infty(D)}\}.$$

To conclude the proof, we show that the first term of the right-hand side of last inequality is of higher order in ε, namely of order ε^3. Indeed, we have

$$\|a-a_0\|_{L^\infty(D)} = \varepsilon\sum_{j=1}^{L}|Y_j|\|a_j\|_{L^\infty(D)} \leq c_1\varepsilon$$

and, taking $v = u - (u_0 + \varepsilon u_1)$ in (25),

$$\|\nabla(u-(u_0+\varepsilon u_1))\|_{L^2(D)} \leq \frac{1}{a_{min}}\varepsilon^2\sum_{i,j=1}^{L}|Y_iY_j|\|a_i\|_{L^\infty(D)}\|\nabla U_j\|_{L^2(D)} \leq c_2\varepsilon^2 \tag{27}$$

with c_1, c_2 two (deterministic) constants independent of u, h and ε. Therefore, we have

$$\mathbb{E}\left[\|a-a_0\|_{L^\infty(D)}\|\nabla(u-(u_0+\varepsilon u_1))\|_{L^2(D)}\right] \leq C_2\varepsilon^3$$

with $C_2 = c_1c_2$. □

Remark 2.6. *A bound for $\|\mathbb{E}[u-u_0]\|_{H_0^1(D)}$ can also be obtained using Jensen's inequality, the fact that the term u_1 is mean-free and (27) as follows*

$$\|\mathbb{E}[u-u_0]\|_{H_0^1(D)} = \|\mathbb{E}[u-u_0-\varepsilon u_1]\|_{H_0^1(D)} \leq \mathbb{E}[\|\nabla(u-(u_0+\varepsilon u_1))\|_{L^2(D)}] \leq \frac{\varepsilon^2\sigma^2 C_P}{a_{min}a^2_{0,min}}\|f\|_{L^2(D)}\left(\sum_{j=1}^{L}\|a_j\|_{L^\infty(D)}\right)^2.$$

Compared to (24), there is no additional higher order term here but the constant for the term of order ε^2 is larger since the cross terms do not vanish and $a^{-1}_{0,min}$ is replaced by a^{-1}_{min}.

2.2.2 A posteriori error estimator in the $L^2_\rho(\Gamma; H^1_0(D))$-norm

The goal is now to have an estimation of the error between u and $u_{0,h}$ which does not depend on the exact solution u. Let us define the jump of a function φ across an edge $e \in \mathcal{T}_h$ in the direction of $\mathbf{n}_e$ by

$$[\varphi]_{\mathbf{n}_e}(x) := \begin{cases} \lim_{t\to 0^+} \left(\varphi(x + t\mathbf{n}_e) - \varphi(x - t\mathbf{n}_e)\right) & \text{if } e \not\subseteq \partial D \\ 0 & \text{if } e \subset \partial D, \end{cases}$$

where $\mathbf{n}_e$ denotes a normal of e of arbitrary (but fixed) direction for internal edges and the outwards normal to ∂D if $e \in \partial D$. Notice that the quantity $[\nabla\varphi \cdot \mathbf{n}_e]_{\mathbf{n}_e}$ is independent of the choice of the direction of the normal $\mathbf{n}_e$. We obtain the following residual type error estimator proceeding as in [33].

Proposition 2.7. *Let u and $u_{0,h}$ be the solution of problems (2) and (18) respectively. There exists a constant $C > 0$ depending only on the constants in (15) and (16) such that*

$$\mathbb{E}\left[\|\nabla(u - u_{0,h})\|^2_{L^2(D)}\right]^{\frac{1}{2}} \leq \frac{\sqrt{2}}{a_{min}} \left[C\eta_1^2 + \eta_2^2\right]^{\frac{1}{2}}, \tag{28}$$

with

$$\eta_1^2 := \sum_{K\in\mathcal{T}_h} h_K^2 \int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))^2 + \sum_{e\in\mathcal{T}_h} h_e \int_e [a_0\nabla u_{0,h}\cdot\mathbf{n_e}]^2_{\mathbf{n_e}}$$

$$\eta_2^2 := \varepsilon^2\sigma^2 \int_D \sum_{j=1}^{L} a_j^2 |\nabla u_{0,h}|^2.$$

Proof. In the sequel, C will denote a constant whose value might change from one line to another. Let v be any function in $H^1_0(D)$. We have

$$\begin{aligned} \int_D a\nabla(u - u_{0,h})\cdot\nabla v &= \int_D a\nabla u\cdot\nabla v - \int_D a\nabla u_{0,h}\cdot\nabla v \\ &= \underbrace{\int_D (fv - a_0\nabla u_{0,h}\cdot\nabla v)}_{:=A_1} + \underbrace{\int_D (a_0 - a)\nabla u_{0,h}\cdot\nabla v}_{:=A_2}, \end{aligned} \tag{29}$$

where A_1 and A_2 correspond respectively to the residual for u_0, solution to problem (10), and the error due to the approximation of u by u_0. We bound now each term separately, starting with A_2. Using the expansion of a given by (3), we have

$$A_2 \leq \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2\right)^{\frac{1}{2}} \left(\int_D |\nabla v|^2\right)^{\frac{1}{2}} = \varepsilon \left(\int_D (\sum_{j=1}^{L} a_j Y_j)^2 |\nabla u_{0,h}|^2\right)^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}. \tag{30}$$

Since $\int_D a_0\nabla u_{0,h}\cdot\nabla v_h = \int_D f v_h$ for all $v_h \in V_h$, we get for the first term A_1

$$\begin{aligned} A_1 &= \sum_{K\in\mathcal{T}_h} \int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))\,(v - v_h) + \sum_{e\in\mathcal{T}_h} \int_e [a_0\nabla u_{0,h}\cdot\mathbf{n_e}]_{\mathbf{n_e}}(v - v_h) \\ &\leq \sum_{K\in\mathcal{T}_h} \left(\int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))^2\right)^{\frac{1}{2}} \left(\int_K (v - v_h)^2\right)^{\frac{1}{2}} \\ &\quad + \sum_{e\in\mathcal{T}_h} \left(\int_e [a_0\nabla u_{0,h}\cdot\mathbf{n_e}]^2_{\mathbf{n_e}}\right)^{\frac{1}{2}} \left(\int_e (v - v_h)^2\right)^{\frac{1}{2}}, \end{aligned}$$

for any $v_h \in V_h$. Thanks to the interpolation results (15) and (16), if we take v_h as being the Clément's interpolation of v, we obtain

$$\begin{aligned}
A_1 &\leq \sum_{K\in\mathcal{T}_h} \left(\int_K |f + \nabla\cdot(a_0\nabla u_{0,h})|^2 \right)^{\frac{1}{2}} C h_K |v|_{H^1(N(K))} + \sum_{e\in\mathcal{T}_h} \left(\int_e [a_0\nabla u_{0,h}\cdot \mathbf{n_e}]^2_{\mathbf{n_e}} \right)^{\frac{1}{2}} C h_e^{\frac{1}{2}} |v|_{H^1(N(K_e))} \\
&\leq C \left[\left(\sum_{K\in\mathcal{T}_h} h_K^2 \int_K |f + \nabla\cdot(a_0\nabla u_{0,h})|^2 \right)^{\frac{1}{2}} + \left(\sum_{e\in\mathcal{T}_h} h_e \int_e [a_0\nabla u_{0,h}\cdot \mathbf{n_e}]^2_{\mathbf{n_e}} \right)^{\frac{1}{2}} \right] \|\nabla v\|_{L^2(D)} \\
&\leq \sqrt{2} C \left[\sum_{K\in\mathcal{T}_h} h_K^2 \int_K |f + \nabla\cdot(a_0\nabla u_{0,h})|^2 + \sum_{e\in\mathcal{T}_h} h_e \int_e [a_0\nabla u_{0,h}\cdot \mathbf{n_e}]^2_{\mathbf{n_e}} \right]^{\frac{1}{2}} \|\nabla v\|_{L^2(D)}. \qquad (31)
\end{aligned}$$

Let us now take $v = u(\cdot,\mathbf{y}) - u_{0,h} \in H_0^1(D)$. Since a_{min} is a lower bound for a, we deduce from (29) that

$$\int_D |\nabla(u - u_{0,h})|^2 \leq \frac{1}{a_{min}} \left[A_1 + A_2 \right].$$

Combining this last inequality with the bounds for A_1 and A_2 given by (31) and (30) respectively, we obtain

$$\begin{aligned}
\|\nabla(u - u_{0,h})\|_{L^2(D)} \leq \frac{1}{a_{min}} \Bigg\{ \sqrt{2} C \Bigg[& \sum_{K\in\mathcal{T}_h} h_K^2 \int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))^2 \\
& + \sum_{e\in\mathcal{T}_h} h_e \int_e [a_0\nabla u_{0,h}\cdot \mathbf{n_e}]^2_{\mathbf{n_e}} \Bigg]^{\frac{1}{2}} + \varepsilon \left(\int_D (\sum_{j=1}^{L} a_j Y_j)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \Bigg\}. \qquad (32)
\end{aligned}$$

Taking the square of this last equation and using again $(a+b)^2 \leq 2(a^2+b^2)$ yields

$$\begin{aligned}
\|\nabla(u - u_{0,h})\|^2_{L^2(D)} \leq \frac{2}{a_{min}^2} \Bigg\{ 2C^2 \Bigg(& \sum_{K\in\mathcal{T}_h} h_K^2 \int_K |f + \nabla\cdot(a_0\nabla u_{0,h})|^2 \\
& + \sum_{e\in\mathcal{T}_h} h_e \int_e [a_0\nabla u_{0,h}\cdot \mathbf{n_e}]^2_{\mathbf{n_e}} \Bigg) + \varepsilon^2 \int_D (\sum_{j=1}^{L} a_j Y_j)^2 |\nabla u_{0,h}|^2 \Bigg\}.
\end{aligned}$$

The *a posteriori* error estimation (28) is obtained taking the square root of the expected value on both sides of the last inequality and exploiting the independence of the random variables, namely that $\mathbb{E}[Y_i Y_j] = \sigma^2 \delta_{ij}$ for $i,j = 1,...,L$ where δ_{ij} denotes the Kronecker delta. □

Remark 2.8. *In the one-dimensional case, we can take $v_h = r_h v$ the Lagrange interpolant of v and the sum over the edges (the discrete nodes here) vanishes. Indeed, any function and its Lagrange interpolant coincide on each node x_i, $i = 0,...,N+1$, of the considered discretization, or more precisely $v(x_i) - r_h v(x_i) = 0$ for all $i = 0,...,N+1$. Since (13) holds for e.g. $C = 2$, we can show that we have the following a posteriori error estimator*

$$\mathbb{E}\left[\|u' - u'_{0,h}\|^2_{L^2(D)} \right]^{\frac{1}{2}} \leq \frac{\sqrt{2}}{a_{min}} \left(4 \sum_{i=0}^{N} h_i^2 \int_{x_i}^{x_{i+1}} (f + (a_0 u'_{0,h})')^2 + \varepsilon^2 \sigma^2 \int_D \sum_{j=1}^{L} a_j^2 (u'_{0,h})^2 \right)^{\frac{1}{2}}, \qquad (33)$$

where u' denotes the spatial derivative $\frac{\partial u(x,\omega)}{\partial x}$.

2.2.3 A posteriori error estimator in the $L^2_\rho(\Gamma; L^2(D))$-norm

We now give an *a posteriori* error estimator of the error between u and $u_{0,h}$ in the L^2-norm in space, which lead to a gain of one order in h. To do so, we use a duality argument (often called *Aubin-Nitsche* trick). We thus consider the dual problem of Problem (2) given by:

find $\phi : D \times \Omega \to \mathbb{R}$ such that P-almost everywhere:

$$\left\{ \begin{array}{rcll} -\operatorname{div}\left(a(\mathbf{x},\omega)\nabla\phi(\mathbf{x},\omega)\right) & = & u(\mathbf{x},\omega) - u_{0,h}(\mathbf{x}) & \text{in } D \\ \phi(\mathbf{x},\omega) & = & 0 & \text{on } \partial D, \end{array} \right. \tag{34}$$

whose point-wise in $\mathbf{y} \in \Gamma$ weak form reads:

for a.e. $\mathbf{y} \in \Gamma$, find $\phi(\cdot, \mathbf{y}) \in H^1_0(D)$ such that

$$\int_D a(\mathbf{x},\mathbf{y})\nabla\phi(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{x}) dx = \int_D (u(\mathbf{x},\mathbf{y}) - u_{0,h}(\mathbf{x}))v(\mathbf{x}) dx \quad \forall v \in H^1_0(D). \tag{35}$$

Under regularity conditions on D, we have the following *a posteriori* error estimator, which implies that the convergence rate of the error is $\mathcal{O}(h^2 + \varepsilon)$ in that case, i.e. that we gain one order in h compared to the error in the norm $L^2_\rho(\Gamma; H^1_0(D))$. However, the order of the statistical error is not improved.

Proposition 2.9. *Let u, u_0 and $u_{0,h}$ be the solution of problems (2), (10) and (18) respectively. If $\phi(\cdot, \mathbf{y}) \in H^2(D)$ and $\|\phi\|_{H^2(D)} \leq C\|u - u_{0,h}\|_{L^2(D)}$ for a.e. $\mathbf{y} \in \Gamma$, then there exist constants $C_1, C_2 > 0$ independent of u, h and ε such that*

$$\mathbb{E}\left[\|u - u_{0,h}\|^2_{L^2(D)}\right]^{\frac{1}{2}} \leq \sqrt{2}\left[C_1\eta_1^2 + C_2\eta_2^2\right]^{\frac{1}{2}} \tag{36}$$

with

$$\eta_1^2 := \sum_{K \in \mathcal{T}_h} h_K^4 \int_K \left(f + \nabla \cdot (a_0 \nabla u_{0,h})\right)^2 + \sum_{e \in \mathcal{T}_h} h_e^3 \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n}_e]^2_{\mathbf{n}_e}$$

$$\eta_2^2 := \varepsilon^2\sigma^2 \int_D \sum_{j=1}^{L} a_j^2 |\nabla u_{0,h}|^2.$$

Remark 2.10. *The assumptions of proposition 2.9 on the regularity of the dual solution ϕ are satisfied if, for instance, D is a convex polygon in $\mathbb{R}^2$.*

Proof. First note that if we take $v = u(\cdot, \mathbf{y}) - u_{0,h}$ in (35), we directly get the L^2-norm in space of the error from the right-hand side. We thus only need to estimate the left-hand side by a quantity which does not depends on the exact solutions u and ϕ of respectively the primal and dual problems. Since

$$\int_D a\nabla(u - u_{0,h}) \cdot \nabla v_h + \int_D (a - a_0)\nabla u_{0,h} \cdot \nabla v_h = 0 \quad \forall v_h \in V_h,$$

we have for any $v_h \in V_h$

$$\begin{aligned} \|u - u_{0,h}\|^2_{L^2(D)} &= \int_D a\nabla(u - u_{0,h}) \cdot \nabla\phi \\ &= \int_D a\nabla(u - u_{0,h}) \cdot \nabla(\phi - v_h) - \int_D (a - a_0)\nabla u_{0,h} \cdot \nabla v_h \\ &= \underbrace{\int_D f(\phi - v_h) - \int_D a_0 \nabla u_{0,h} \nabla(\phi - v_h)}_{:=A_1} \underbrace{- \int_D (a - a_0)\nabla u_{0,h} \cdot \nabla\phi}_{:=A_2}. \end{aligned} \tag{37}$$

We now treat each term separately. For the first one, we follow the usual procedure. For any $v_h \in V_h$, we have

$$\begin{aligned} A_1 &= \sum_{K\in\mathcal{T}_h} \int_K f(\phi - v_h) - \sum_{K\in\mathcal{T}_h} \int_K a_0 \nabla(\phi - v_h)\nabla u_{0,h} \\ &\leq \sum_{K\in\mathcal{T}_h} \|f + \nabla\cdot(a_0\nabla u_{0,h})\|_{L^2(K)} \|\phi - v_h\|_{L^2(K)} + \sum_{e\in\mathcal{T}_h} \| \left[a_0\nabla u_{0,h}\cdot\mathbf{n}_e\right]_{\mathbf{n}_e} \|_{L^2(e)} \|\phi - v_h\|_{L^2(e)}. \end{aligned}$$

If we take $v_h = r_h\phi$, the Lagrange interpolation of ϕ, thanks to the interpolation error estimate (14), the trace inequality and the standard elliptic regularity result $\|\phi\|_{H^2(D)} \leq C\|u - u_{0,h}\|_{L^2(D)}$ (see [30, 31] for instance), we obtain

$$\begin{aligned} A_1 &\leq C_1 \left[\left(\sum_{K\in\mathcal{T}_h} h_K^4 \int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))^2 \right)^{\frac{1}{2}} + \left(\sum_{e\in\mathcal{T}_h} h_e^3 \int_e [a_0\nabla u_{0,h}\cdot\mathbf{n}_e]^2_{\mathbf{n}_e} \right)^{\frac{1}{2}} \right] |\phi|_{H^2(D)} \\ &\leq \sqrt{2}C_1 \left(\sum_{K\in\mathcal{T}_h} h_K^4 \int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))^2 + \sum_{e\in\mathcal{T}_h} h_e^3 \int_e [a_0\nabla u_{0,h}\cdot\mathbf{n}_e]^2_{\mathbf{n}_e} \right)^{\frac{1}{2}} \|u - u_{0,h}\|_{L^2(D)}, \end{aligned} \tag{38}$$

where C_1 is a constant whose value might change from one line to another. Consider now the second term A_2 of (37). We have

$$A_2 = -\int_D (a - a_0)\nabla u_{0,h}\cdot\nabla\phi \leq \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \|\nabla\phi\|_{L^2(D)},$$

and thus, it only remains to give an estimation of $\|\nabla\phi\|_{L^2(D)}$. Taking $v = \phi$ in the weak form (35) of the dual problem yields

$$\int_D a\nabla\phi\cdot\nabla\phi = \int_D (u - u_{0,h})\phi \leq \|u - u_{0,h}\|_{L^2(D)}\|\phi\|_{L^2(D)}.$$

Since a is bounded from below by a_{min}, thanks to Poincaré inequality we get

$$a_{min}\|\nabla\phi\|^2_{L^2(D)} \leq C_P\|u - u_{0,h}\|_{L^2(D)}\|\nabla\phi\|_{L^2(D)},$$

and thus

$$\|\nabla\phi\|_{L^2(D)} \leq \frac{C_P}{a_{min}}\|u - u_{0,h}\|_{L^2(D)}.$$

Therefore, A_2 can be bounded by

$$A_2 \leq \frac{C_P}{a_{min}} \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}} \|u - u_{0,h}\|_{L^2(D)}. \tag{39}$$

Inserting (38) and (39) into (37) yields

$$\begin{aligned} \|u - u_{0,h}\|_{L^2(D)} \leq \sqrt{2}C_1 &\left(\sum_{K\in\mathcal{T}_h} h_K^4 \int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))^2 + \sum_{e\in\mathcal{T}_h} h_e^3 \int_e [a_0\nabla u_{0,h}\cdot\mathbf{n}_e]^2_{\mathbf{n}_e} \right)^{\frac{1}{2}} \\ &+ \frac{C_P}{a_{min}} \left(\int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right)^{\frac{1}{2}}, \end{aligned}$$

and thus

$$\begin{aligned} \|u - u_{0,h}\|^2_{L^2(D)} \leq 2 &\left[2C_1^2 \left(\sum_{K\in\mathcal{T}_h} h_K^4 \int_K (f + \nabla\cdot(a_0\nabla u_{0,h}))^2 + \sum_{e\in\mathcal{T}_h} h_e^3 \int_e [a_0\nabla u_{0,h}\cdot\mathbf{n}_e]^2_{\mathbf{n}_e} \right) \right. \\ &\left. + \frac{C_P^2}{a_{min}^2} \int_D (a - a_0)^2 |\nabla u_{0,h}|^2 \right]. \end{aligned} \tag{40}$$

Since $\mathbb{E}[(a-a_0)^2] = \varepsilon^2\sigma^2\sum_{j=1}^{L} a_j^2$, the result follows from taking first the expected value and then the square root on both sides of (40). □

2.2.4 Goal-oriented estimator

The *a posteriori* error estimators obtained so far yield bounds on the error in global norms. In the case where we are interested in a particular quantity of interest, e.g. point values or contour integrals, these estimators may not be appropriate. Goal-oriented error estimators have thus been developed (see [34–36] and [18–21] and the references therein for the deterministic and stochastic framework, respectively) to bound a given functional using optimal control techniques (based on a duality-argument). In this section we only sketch the derivation of a goal-oriented estimator for the first-order FEM approximation $u_{0,h}$. Assume that we are interested in computing $Q(u)$ with Q a functional representing a linear quantity of interest which depends on the random vector $\mathbf{y}$ only through the random solution $u(\cdot,\mathbf{y})$ itself. We introduce the dual problem:

$$\text{find } \varphi(\cdot,\mathbf{y}) \in H_0^1(D) \text{ such that } \mathcal{A}(v,\varphi(\cdot,\mathbf{y});\mathbf{y}) = Q(v) \quad \forall v \in H_0^1(D), \text{ a.e. } \mathbf{y}\in\Gamma, \tag{41}$$

where $\mathcal{A}$ is defined by (6). Let $\mathbf{y}_0 = \mathbf{0}$ denote the nominal value for $\mathbf{y}$, i.e. for which $a(\mathbf{x},\mathbf{y}_0) = a_0(\mathbf{x})$ and let φ_0 be the deterministic solution of (41) with $\mathbf{y}=\mathbf{y}_0$ and $\varphi_{0,h}$ its FE approximation. Using the fact that Q does not depend on $\mathbf{y}$ explicitly, we can easily show that for a.e. $\mathbf{y}\in\Gamma$ we have

$$\begin{aligned} Q(u(\cdot,\mathbf{y})) - Q(u_{0,h}) = &\underbrace{\int_D f\varphi_0 - \int_D a_0\nabla u_{0,h}\cdot\nabla\varphi_0}_{A_1} \underbrace{- \int_D (a-a_0)\nabla u_{0,h}\cdot\nabla\varphi_{0,h}}_{A_2} \\ &\underbrace{- \int_D (a-a_0)\nabla u_{0,h}\cdot\nabla(\varphi_0-\varphi_{0,h})}_{A_3} \underbrace{- \int_D (a-a_0)\nabla(u-u_{0,h})\cdot\nabla\varphi_{0,h}}_{A_4} \\ &\underbrace{- \int_D (a-a_0)\nabla(u-u_{0,h})\cdot\nabla(\varphi_0-\varphi_{0,h})}_{A_5}. \end{aligned}$$

The first term A_1, which is deterministic and of order h^2, can be bounded using standard techniques such as the Dual-weighted residual (DWR) method (see e.g. [34, 35]) or using the parallelogram identity as proposed by Oden and Prudhomme in [36]. In the DWR method, the estimator depends on the unknown influence function φ_0, either through $|\varphi_0|_{H^2(K)}$ or $\|\nabla(\varphi_0-\varphi_{0,h})\|_{L^2(K)}$, K being an element of the mesh. In the former case, the H^2 seminorm can be estimated by a discrete analogue and in the latter case, the influence function might be replaced by a discrete solution computed on a space richer than V_h or by post-processing. All the other terms can be bounded provided that we have an estimation of $\|\nabla(u-u_{0,h})\|_{L^2(D)}$, which is given by (32), as well as an estimation of $\|\nabla(\varphi_0-\varphi_{0,h})\|_{L^2(D)}$ which can be done as in the previous sections. Moreover, based on the results obtained in the previous sections we have

$$A_1 = \mathcal{O}(h^2), A_2 = \mathcal{O}(\varepsilon), A_3 = \mathcal{O}(h\varepsilon), A_4 = \mathcal{O}(h\varepsilon+\varepsilon^2) \text{ and } A_5 = \mathcal{O}(h^2\varepsilon+\varepsilon^2 h).$$

Usually, we are interested in computing the expectation or the variance of $Q(u)-Q(u_{0,h})$. In the former case, notice that $\mathbb{E}[A_2] = \mathbb{E}[A_3] = 0$ and since A_1 is a deterministic quantity, we have

$$\mathbb{E}[Q(u)-Q(u_{0,h})] = A_1 + \mathbb{E}[A_4] + \mathbb{E}[A_5].$$

Moreover, the term $\mathbb{E}[A_5]$ is of higher order than $\mathbb{E}[A_4]$ and can thus be neglected, so that we have $\mathbb{E}[Q(u)-Q(u_{0,h})] = \mathcal{O}(h^2+h\varepsilon+\varepsilon^2)$. In the latter case, we have

$$\mathbb{E}[|Q(u)-Q(u_{0,h})|^2] \le 5\left(A_1^2 + \mathbb{E}[A_2^2] + \mathbb{E}[A_3^2] + \mathbb{E}[A_4^2] + \mathbb{E}[A_5^2]\right).$$

As before, the term $\mathbb{E}[A_5^2]$ can be neglected and we have $\mathbb{E}[|Q(u)-Q(u_{0,h})|^2]^{\frac{1}{2}} = \mathcal{O}(h^2+\varepsilon+h\varepsilon)$. Moreover, if the mesh space h is chosen such that $h^2 \sim \varepsilon$, then both terms $\mathbb{E}[A_3^2]$ and $\mathbb{E}[A_4^2]$ can also be omitted in the estimation of the variance and $\mathbb{E}[|Q(u)-Q(u_{0,h})|^2]^{\frac{1}{2}} = \mathcal{O}(h^2+\varepsilon)$.

2.3 Second order approximation

In this section, instead of considering the error between u and $u_{0,h}$, we will give an estimation of the error between u and u_h^1, the FE approximation of $u^1 := u_0 + \varepsilon u_1 = u_0 + \varepsilon \sum_{j=1}^{L} U_j Y_j$, where U_j is the solution of problem (11). Since the random variables Y_j, $j = 1, ..., L$, are assumed to be bounded, the error due to the stochastic approximation of u is of order ε^2 in this case. Indeed, if we don't take the finite element approximation error into account, proceeding as in (17) we have

$$\begin{aligned}\int_D a\nabla(u-u^1)\cdot\nabla v = & -\varepsilon\int_D a_0\nabla u_1\cdot\nabla v - \int_D (a-a_0)\nabla u^1\cdot\nabla v \\ = & -\varepsilon\sum_{j=1}^{L} Y_j\int_D (a_0\nabla U_j + a_j\nabla u_0)\cdot\nabla v - \varepsilon^2\int_D\sum_{i,j=1}^{L} Y_iY_j a_j\nabla U_i\cdot\nabla v \\ = & -\varepsilon^2\int_D\sum_{i,j=1}^{L} Y_iY_j a_j\nabla U_i\cdot\nabla v, \end{aligned} \tag{42}$$

and only the term of order ε^2 remains. Let us now take the error due to the approximation of u^1 by $u_h^1 := u_{0,h} + \varepsilon u_{1,h}$ into account, where $u_{1,h} = \sum_{j=1}^{L} Y_j U_{j,h}$ and, for $j = 1, ..., L$, $U_{j,h}$ is the solution of

$$\int_D a_0\nabla U_{j,h}\cdot\nabla v_h = -\int_D a_j\nabla u_{0,h}\cdot\nabla v_h \quad \forall v_h \in V_h. \tag{43}$$

To simplify the notation, we define

$$w_{j,h} := a_0\nabla U_{j,h} + a_j\nabla u_{0,h}.$$

We can show that convergence of the error is in $\mathcal{O}(h+\varepsilon h+\varepsilon^2)$, i.e., that for a mesh size h of order ε^2, the error is divided by 4 when ε is halved. The following proposition provides an *a posteriori* error estimator.

Proposition 2.11. *Let u, $u_{0,h}$ and $U_{j,h}$, $j = 1, ..., L$, be the solutions of Problems (2), (18) and (43) respectively. There exist two constants $C_1, C_2 > 0$ depending only on the constants in (15) and (16) such that*

$$\mathbb{E}\left[\|\nabla(u-u_h^1)\|_{L^2(D)}^2\right]^{\frac{1}{2}} \leq \frac{\sqrt{3}}{a_{min}}\left[C_1\eta_1^2 + C_2\eta_2^2 + \eta_3^2\right]^{\frac{1}{2}}, \tag{44}$$

with

$$\begin{aligned}\eta_1^2 &= \sum_K h_K^2\|f + \nabla\cdot(a_0\nabla u_{0,h})\|_{L^2(K)}^2 + \sum_e h_e\|\left[a_0\nabla u_{0,h}\cdot\mathbf{n}_e\right]_{\mathbf{n}_e}\|_{L^2(e)}^2, \\ \eta_2^2 &= \varepsilon^2\sigma^2\left(\sum_K h_K^2\int_K\sum_{j=1}^{L}(\nabla\cdot w_{j,h})^2 + \sum_e h_e\int_e\sum_{j=1}^{L}[w_{j,h}\cdot\mathbf{n}_e]_{\mathbf{n}_e}^2\right), \\ \eta_3^2 &= \varepsilon^4\left(\int_D\sum_{i=1}^{L} a_i^2|\nabla U_{i,h}|^2\mathbb{E}[Y_i^4] + \sigma^4\int_D\sum_{\substack{i,j=1\\ i\neq j}}^{L}\left[a_i^2|\nabla U_{j,h}|^2 + 2a_ia_j\nabla U_{i,h}\cdot\nabla U_{j,h}\right]\right).\end{aligned}$$

From (44), we see that the error splits into three parts, namely the error due to the FE approximation of u_0, the FE approximation of the U_j, $j = 1, ..., L$ and the truncation in the expansion of u with respect to ε.

Proof. For any $v \in H_0^1(D)$, we have

$$\int_D a\nabla(u-u_h^1)\cdot\nabla v = \underbrace{\int_D fv - \int_D a_0\nabla u_{0,h}\cdot\nabla v}_{:=A_1} \underbrace{-\varepsilon\int_D \sum_{j=1}^{L} Y_j(a_0\nabla U_{j,h} + a_j\nabla u_{0,h})\cdot\nabla v}_{:=A_2}$$
$$\underbrace{-\varepsilon\int_D (a-a_0)\nabla u_{1,h}\cdot\nabla v}_{:=A_3}. \tag{45}$$

where A_1 is again the residual for u_0, while A_2 and A_3 correspond respectively to the error due to the approximation of U_j by $U_{j,h}$, for $j = 1, ..., L$ and the approximation of u by $u_0 + \varepsilon u_1$. Let us treat each term separately. The first term A_1 is bounded by (see section 2.2)

$$A_1 \le C_1\left[\sum_{K\in\mathcal{T}_h} h_K^2\|f+\nabla\cdot(a_0\nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e\in\mathcal{T}_h} h_e\|\left[a_0\nabla u_{0,h}\cdot\mathbf{n}_e\right]_{\mathbf{n}_e}\|_{L^2(e)}^2\right]^{\frac{1}{2}}\|\nabla v\|_{L^2(D)}. \tag{46}$$

Let us consider now the term A_2. Since $\int_D w_{j,h}\cdot\nabla v_h = 0$ for all $v_h \in V_h$, we have

$$\begin{aligned} A_2 =& -\varepsilon\int_D \sum_{j=1}^{L} Y_j w_{j,h}\cdot\nabla(v-\mathcal{I}_h v) \\ =& \varepsilon\sum_{K\in\mathcal{T}_h}\int_K (\sum_{j=1}^{L} Y_j\nabla\cdot w_{j,h})(v-\mathcal{I}_h v) + \varepsilon\sum_{e\in\mathcal{T}_h}\int_e [\sum_{j=1}^{L} Y_j w_{j,h}\cdot\mathbf{n}_e]_{\mathbf{n}_e}(v-\mathcal{I}_h v) \\ \le& C_2\left(\sum_{K\in\mathcal{T}_h}\varepsilon^2 h_K^2\|\sum_{j=1}^{L} Y_j\nabla\cdot w_{j,h}\|_{L^2(K)}^2 + \sum_{e\in\mathcal{T}_h}\varepsilon^2 h_e\|[\sum_{j=1}^{L} Y_j w_{j,h}\cdot\mathbf{n}_e]_{\mathbf{n}_e}\|_{L^2(e)}^2\right)^{\frac{1}{2}}\|\nabla v\|_{L^2(D)}, \end{aligned} \tag{47}$$

where C_2 depends only on the interpolation constants that appear in (15) and (16). Finally, we estimate the last term A_3. We have

$$\begin{aligned} A_3 =& -\varepsilon\int_D (\varepsilon\sum_{j=1}^{L} Y_j a_j)\nabla(\sum_{i=1}^{L} Y_i U_{i,h})\cdot\nabla v = -\varepsilon^2\int_D \sum_{i,j=1}^{L} Y_i Y_j a_j\nabla U_{i,h}\cdot\nabla v \\ \le& \varepsilon^2\|\sum_{i,j=1}^{L} Y_i Y_j a_j\nabla U_{i,h}\|_{L^2(D)}\|\nabla v\|_{L^2(D)}. \end{aligned} \tag{48}$$

Since a is bounded from below by a_{min}, combining (45) with (46), (47) and (48) with $v = u(\cdot,\mathbf{y}) - u_h^1(\cdot,\mathbf{y}) \in H_0^1(D)$ yields

$$\begin{aligned} &\|\nabla(u-u_h^1)\|_{L^2(D)} \le \\ &\frac{\sqrt{3}}{a_{min}}\left[C_1^2\left(\sum_{K\in\mathcal{T}_h} h_K^2\|f+\nabla\cdot(a_0\nabla u_{0,h})\|_{L^2(K)}^2 + \sum_{e\in\mathcal{T}_h} h_e\|\left[a_0\nabla u_{0,h}\cdot\mathbf{n}_e\right]_{\mathbf{n}_e}\|_{L^2(e)}^2\right)\right. \\ &+ C_2^2\left(\sum_{K\in\mathcal{T}_h}\varepsilon^2 h_K^2\|\sum_{j=1}^{L} Y_j\nabla\cdot w_{j,h}\|_{L^2(K)}^2 + \sum_{e\in\mathcal{T}_h}\varepsilon^2 h_e\|[\sum_{j=1}^{L} Y_j w_{j,h}\cdot\mathbf{n}_e]_{\mathbf{n}_e}\|_{L^2(e)}^2\right) \\ &\left.+ \varepsilon^4\|\sum_{i,j=1}^{L} Y_i Y_j a_j\nabla U_{i,h}\|_{L^2(D)}^2\right]^{\frac{1}{2}}, \end{aligned}$$

using the inequality $(a+b+c) \leq \sqrt{3}(a^2+b^2+c^2)^{\frac{1}{2}}$. To conclude the proof, it only remains to take the expected value on both sides of the square of this last inequality. By linearity of the expected value, we can consider the three terms of the right-hand side separately. The first term is a deterministic quantity and thus, taking the expected value on it has no effect. For the two other terms, we just have to evaluate $\mathbb{E}[Y_iY_j]$ for $1 \leq i,j \leq L$ and $\mathbb{E}[Y_iY_jY_kY_l]$ for $1 \leq i,j,k,l \leq L$. Since the random variables are supposed to be independent, of zero-mean and finite variance σ^2, we have $\mathbb{E}[Y_iY_j] = \sigma^2\delta_{ij}$. Furthermore, we have

$$\mathbb{E}[Y_iY_jY_kY_l] = \begin{cases} \mathbb{E}[Y_j^4] & \text{if } i=j=k=l \\ \sigma^4 & \text{if the indices are pairwise equal} \\ 0 & \text{otherwise.} \end{cases}$$

Let us write

$$B := \sum_{i,j,k,l=1}^{L} Y_iY_jY_kY_l a_j a_k \nabla U_{i,h} \cdot \nabla U_{l,h},$$

which we split into three parts B_1 (all indices are equal), B_2 (two pairs of indices) and B_3 (remaining indices). Thanks to the linearity of expectation, we have $\mathbb{E}[B] = \mathbb{E}[B_1] + \mathbb{E}[B_2] + \mathbb{E}[B_3]$. First, we can notice that $\mathbb{E}[B_3] = 0$. Moreover, the contribution to $\mathbb{E}[B]$ when $i=j=k=l$ is

$$\mathbb{E}[B_1] = \sum_{i=1}^{L} a_i^2 |\nabla U_{i,h}|^2 \mathbb{E}[Y_i^4].$$

Let us consider now all the cases when we have pairwise equal pairs of indices. Out of 4 indices, there are three different ways to form two pairs of indices, namely $(j=k, i=l)$, $(j=i, k=l)$ and $(j=l, k=i)$. Since the two last cases lead to the same result, we get

$$\mathbb{E}[B_2] = \sigma^4 \left(\sum_{\substack{i,j=1 \\ i \neq j}}^{L} a_j^2 |\nabla U_{i,h}|^2 + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^{L} a_i a_j \nabla U_{i,h} \cdot \nabla U_{j,h} \right).$$

Altogether, we finally get

$$\mathbb{E}[B] = \sum_{i=1}^{L} a_i^2 |\nabla U_{i,h}|^2 \mathbb{E}[Y_i^4] + \sigma^4 \sum_{\substack{i,j=1 \\ i \neq j}}^{L} \left[a_i^2 |\nabla U_{j,h}|^2 + 2 a_i a_j \nabla U_{i,h} \cdot \nabla U_{j,h} \right],$$

which concludes the proof. □

2.4 Generalization

Suppose now that the random solution u of problem (2) is expanded with respect to ε up to order $N \in \mathbb{N}$, see (9). For $1 \leq n \leq N$, let us write

$$u_n(\mathbf{x}, \mathbf{y}(\omega)) = \sum_{j_1, j_2, \ldots, j_n = 1}^{L} U_{j_1 j_2 \cdots j_n}(\mathbf{x}) Y_{j_1}(\omega) Y_{j_2}(\omega) \cdots Y_{j_n}(\omega) \tag{49}$$

the n^{th} term in the expansion. The L^n functions $U_{j_1 j_2 \cdots j_n}$ are obtained by solving for $j_1, j_2, ..., j_n = 1, ..., L$ the deterministic problem

$$\begin{cases} -\operatorname{div}\left(a_{j_1}(\mathbf{x}) \nabla U_{j_2 \cdots j_n}(\mathbf{x}) + a_0(\mathbf{x}) \nabla U_{j_1 \cdots j_n}(\mathbf{x})\right) &= 0, \quad \mathbf{x} \in D \\ U_{j_1 \cdots j_n}(\mathbf{x}) &= 0, \quad \mathbf{x} \in \partial D \end{cases} \tag{50}$$

using the solutions $U_{j_2\cdots j_n}$, $j_2,...,j_n = 1,...,L$, obtained for the $(n-1)^{th}$ term. Proceeding as in Sections 2.2 and 2.3, it is easy to show that the error due to the truncation in the expansion of u is of order ε^{N+1}. More precisely, we have for any $v \in H_0^1(D)$ and almost surely

$$\int_D a\nabla\left(u - \sum_{n=0}^{N}\varepsilon^n u_n\right)\cdot\nabla v = -\varepsilon^{N+1}\sum_{j_0,j_1,...,j_N=1}^{L} Y_{j_0}Y_{j_1}\cdots Y_{j_N}\int_D a_{j_0}\nabla U_{j_1 j_2\cdots j_N}\cdot\nabla v. \tag{51}$$

Since Y_j, $j = 1,...,L$ are bounded, in particular they have bounded $2(N+1)^{th}$ moment. When the various deterministic functions are approximated using finite elements, we can show that the error $u - \sum_{n=0}^{N}\varepsilon^n u_{n,h}$ in the $L^2_\rho(\Gamma; H_0^1(D))$ norm is of order

$$h + \varepsilon h + \varepsilon^2 h + ... + \varepsilon^N h + \varepsilon^{N+1}.$$

The error in $\mathcal{O}(\varepsilon^n h)$, $0 \le n \le N$, corresponds to the error made when the functions $U_{j_1\cdots j_n}$ (u_0 for n=0) are replaced by their FE approximation $U_{j_1\cdots j_n,h}$ (resp. $u_{0,h}$). An *a posteriori* error estimator can thus easily be obtained as follows. First, the term in $\mathcal{O}(h)$, which corresponds to the residual for u_0, is obtained by estimating $\int_D (fv - a_0\nabla u_{0,h}\cdot\nabla v)$, see (31). For the term in $\mathcal{O}(h\varepsilon^n)$, $n = 1,...,N$, it suffices to estimate for $j_1,...,j_n = 1,...,L$ the residual defined for any $v \in H_0^1(D)$ by

$$\langle \mathcal{R}(U_{j_1\cdots j_n,h}), v\rangle := \int_D (a_{j_1}\nabla U_{j_2\cdots j_n,h} + a_0\nabla U_{j_1\cdots j_n,h})\cdot\nabla v,$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing bracket. For an explicit estimator, we finally need to express the expectation of the product of n random variables $\mathbb{E}[Y_{j_1}\cdots Y_{j_n}]$ for all combinations of indices and for $n = 1,...,2(N+1)$. More precisely, we can show the following result.

Proposition 2.12. *Let u be the solution of problem (2) and $u_h^N = \sum_{n=0}^{N}\varepsilon^n u_{n,h}$, where $u_{n,h}$ is the FE approximation of u_n given by (49). There exist $N+1$ constants $C_n > 0$, $n = 0,1,...,N$, depending only on the constants in (15) and (16) such that*

$$\mathbb{E}\left[\|\nabla(u - u_h^N)\|^2_{L^2(D)}\right]^{\frac{1}{2}} \le \frac{\sqrt{N+2}}{a_{min}}\left[C_0\eta_0^2 + \sum_{n=1}^{N} C_n\eta_n^2 + \eta_{N+1}^2\right]^{\frac{1}{2}}, \tag{52}$$

with

$$\eta_0^2 = \sum_K h_K^2\|f + \nabla\cdot(a_0\nabla u_{0,h})\|^2_{L^2(K)} + \sum_e h_e\|\,[a_0\nabla u_{0,h}\cdot\mathbf{n}_e]_{\mathbf{n}_e}\,\|^2_{L^2(e)},$$

$$\eta_n^2 = \varepsilon^{2n}\mathbb{E}\left[\sum_K h_K^2\|\sum_{j_1,...,j_n=1}^{L} Y_{j_1}\cdots Y_{j_n}\nabla\cdot w_{j_1\cdots j_n,h}\|^2_{L^2(K)}\right.$$
$$\left. + \sum_e h_e\|[\sum_{j_1,...,j_n=1}^{L} Y_{j_1}\cdots Y_{j_n} w_{j_1\cdots j_n,h}\cdot\mathbf{n}_e]_{\mathbf{n}_e}\|^2_{L^2(e)}\right]$$

$$\eta_{N+1}^2 = \varepsilon^{2(N+1)}\mathbb{E}\left[\|\sum_{j_0,j_1,...,j_N=1}^{L} Y_{j_0}Y_{j_1}\cdots Y_{j_N}a_{j_0}\nabla U_{j_1\cdots j_N,h}\|^2_{L^2(D)}\right],$$

where

$$w_{j_1\cdots j_n,h} := a_{j_1}\nabla U_{j_2\cdots j_n,h} + a_0\nabla U_{j_1\cdots j_n,h} \quad j_1,...,j_n = 1,...,L.$$

3 Nonlinear problems

Keeping the same notations as in the previous sections, we are now interested in solving problems of the form:

find $u : D \times \Omega \to \mathbb{R}$ such that almost surely:

$$\left\{ \begin{array}{rcll} F(a,u) & = & 0 & \text{in } D \\ u & = & 0 & \text{on } \partial D, \end{array} \right. \tag{53}$$

where F is a smooth nonlinear mapping that depends on the uncertain input a given by (3). Again, the random solution u is expanded with respect to ε up to a certain order

$$u(\mathbf{x}, \mathbf{y}(\omega)) = u_0(\mathbf{x}) + \varepsilon u_1(\mathbf{x}, \mathbf{y}(\omega)) + \mathcal{O}(\varepsilon^2).$$

Formally, we have

$$F(a,u) = F(a_0,u_0) + D_a F(a_0,u_0)(a-a_0) + D_u F(a_0,u_0)(u-u_0) + \mathcal{O}(\varepsilon^2),$$

where D_a and D_u denote the Fréchet derivative with respect to a and u respectively, the deterministic part u_0 of u is the solution of the (nonlinear) problem

$$\left\{ \begin{array}{rcll} F(a_0,u_0) & = & 0 & \text{in } D \\ u_0 & = & 0 & \text{on } \partial D, \end{array} \right. \tag{54}$$

while the U_j in $u_1 = \sum_{j=1}^{L} Y_j U_j$ can be found by solving the (linear) problems

$$\left\{ \begin{array}{rcll} D_a F(a_0,u_0)(a_j) + D_u F(a_0,u_0)(U_j) & = & 0 & \text{in } D \\ U_j & = & 0 & \text{on } \partial D, \end{array} \right. \quad j = 1, ..., L. \tag{55}$$

We can directly see one of the advantages of expanding the solution as proposed here, namely that a single nonlinear problem must be solved to find u_0, the other problems being linear. A new FE solver corresponding to (55) has to be implemented to approximate the U_j, $j = 1, ..., L$.

In the case of quasi-linear problems, the error analysis is very similar to the linear case considered in Section 2. Indeed, under certain conditions such as well-posedness of the problem, only the part of the estimator corresponding to the residual error in the physical space has to be changed in the *a posteriori* estimator of the error between u and $u_{0,h}$ in the $L^2_\rho(\Gamma; H^1_0(D))$-norm. For instance, let us consider problem (53) with

$$F(a,u) := -\operatorname{div}(a(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)) + u^3(\mathbf{x},\omega) - f(\mathbf{x}), \tag{56}$$

which is well-posed as problem (2). In this case, we can show the following *a posteriori* error estimator for $\|u - u_{0,h}\|_{L^2_\rho(\Gamma; H^1_0(D))}$, where $u_{0,h} \in V_h$ is the deterministic solution of

$$\int_D a_0 \nabla u_{0,h} \cdot \nabla v_h + \int_D u_{0,h}^3 v_h = \int_D f v_h \quad \forall v_h \in V_h. \tag{57}$$

Proposition 3.1. *Let u be the solution of problem (53) with F given by (56), and let $u_{0,h}$ be the solution of (57). There exists a constant $C > 0$ depending only on the constants in (15) and (16) such that*

$$\mathbb{E}\left[\|\nabla(u - u_{0,h})\|^2_{L^2(D)} \right]^{\frac{1}{2}} \leq \frac{C}{a_{min}} \left[\eta_1^2 + \eta_2^2 \right]^{\frac{1}{2}},$$

with

$$\eta_1^2 := \sum_{K \in \mathcal{T}_h} h_K^2 \int_K (f - u_{0,h}^3 + \nabla \cdot (a_0 \nabla u_{0,h}))^2 + \sum_{e \in \mathcal{T}_h} h_e \int_e [a_0 \nabla u_{0,h} \cdot \mathbf{n_e}]^2_{\mathbf{n_e}}$$

$$\eta_2^2 := \varepsilon^2 \sigma^2 \int_D \sum_{j=1}^{L} a_j^2 |\nabla u_{0,h}|^2.$$

Proof. Since the proof is very similar to the one of Proposition 2.7, we only give the key ingredients here. First, for any $v \in V$ we have almost surely

$$\int_D a\nabla(u-u_{0,h})\cdot\nabla v = \int_D (f-u_{0,h}^3)v - \int_D a_0\nabla u_{0,h}\cdot\nabla v - \int_D (a-a_0)\nabla u_{0,h}\cdot\nabla v - \int_D (u^3-u_{0,h}^3)v.$$

Then, for $v = u - u_{0,h}$ the last term of last equality is non-positive. Indeed, using that

$$u^3 - u_{0,h}^3 = \int_0^1 3(u_{0,h}+t(u-u_{0,h}))^2(u-u_{0,h})dt,$$

we get

$$-\int_D (u^3-u_{0,h}^3)(u-u_{0,h}) = -\int_D\int_0^1 3(u_{0,h}+t(u-u_{0,h}))^2(u-u_{0,h})^2 \leq 0.$$

Therefore, this term can be omitted since we are looking for an upper bound of the error. □

Another example is the following. Let $k > 0$ be such that $\frac{kC_P^2}{a_{min}} < 1$, or in other words $\frac{kC_P^2}{a_{min}} \leq 1-\delta$ for any $\delta \in (0,1)$. If we take

$$F(a,u) := -\operatorname{div}(a(\mathbf{x},\omega)\nabla u(\mathbf{x},\omega)) - g(u(\mathbf{x},\omega)) \tag{58}$$

in problem (53), where g is a Lipschitz function with Lipschitz constant $k_0 \leq k$, then we can show the well-posedness of the problem and the following *a posteriori* error estimator for the error $u - u_{0,h}$, where $u_{0,h} \in V_h$ is the deterministic solution of

$$\int_D a_0\nabla u_{0,h}\cdot\nabla v_h = \int_D g(u_{0,h})v_h \quad \forall v_h \in V_h. \tag{59}$$

Proposition 3.2. *Let u be the solution of problem (53) with F given by (58), and let $u_{0,h}$ be the solution of (59). There exists a constant $C > 0$, depending only on δ and the constants in (15) and (16), such that*

$$\mathbb{E}\left[\|\nabla(u-u_{0,h})\|_{L^2(D)}^2\right]^{\frac{1}{2}} \leq \frac{C}{a_{min}}\left[\eta_1^2+\eta_2^2\right]^{\frac{1}{2}},$$

with

$$\eta_1^2 := \sum_{K\in\mathcal{T}_h} h_K^2\int_K (g(u_{0,h})+\nabla\cdot(a_0\nabla u_{0,h}))^2 + \sum_{e\in\mathcal{T}_h} h_e\int_e [a_0\nabla u_{0,h}\cdot\mathbf{n_e}]_{\mathbf{n_e}}^2$$

$$\eta_2^2 := \varepsilon^2\sigma^2\int_D\sum_{j=1}^L a_j^2|\nabla u_{0,h}|^2.$$

Proof. Again, we only give the key ingredients of the proof. First, for any $v \in V$ we have almost surely

$$\int_D a\nabla(u-u_{0,h})\cdot\nabla v = \underbrace{\int_D g(u_{0,h})v - \int_D a_0\nabla u_{0,h}\cdot\nabla v - \int_D (a-a_0)\nabla u_{0,h}\cdot\nabla v}_{:=A(v)} - \int_D (g(u)-g(u_{0,h}))v. \tag{60}$$

With $v = u - u_{0,h}$, the last term is bounded by

$$-\int_D (g(u)-g(u_{0,h}))(u-u_{0,h}) \leq k_0C_P^2\|\nabla(u-u_{0,h})\|_{L^2(D)}^2. \tag{61}$$

Since

$$a_{min}\|\nabla(u-u_{0,h})\|^2_{L^2(D)} \leq \int_D a|\nabla(u-u_{0,h})|^2,$$

taking (61) to the left-hand side of (60) and using $k_0 C_P^2 \leq a_{min}(1-\delta)$ yield

$$a_{min}\delta\|\nabla(u-u_{0,h})\|^2_{L^2(D)} \leq A(u-u_{0,h}).$$

A bound on $A(u-u_{0,h})$, which contains the residual for u_0 and a term of order ε, is found proceeding exactly as in the proof of Proposition 2.7. □

The constant C that appears in the estimator of Proposition 3.2 is of order δ^{-1}, and thus explodes when δ tends to zero, i.e. when $\frac{k_0 C_P^2}{a_{min}}$ is close to one. In practise, it is usual to restrict to Lipschitz function with Lipschitz constant $k_0 \leq k$ with k such that $k \leq \frac{a_{min}}{2C_P^2}$, so that $\delta \geq \frac{1}{2}$.

Finally, let us consider an example where the uncertain coefficient is associated to the nonlinear term, namely the problem (53) with

$$F(a,u) = -\Delta u(\mathbf{x},\omega) + au^3(\mathbf{x},\omega) - f(\mathbf{x}). \tag{62}$$

In this case, we can show the well-posedness of the problem and the following *a posteriori* error estimator in $H_0^1(D)$-norm in physical space for the first order approximation $u \approx u_{0,h}$, where $u_{0,h}$ is the solution of

$$\int_D \nabla u_{0,h}\cdot\nabla v_h + \int_D a_0 u_{0,h}^3 v_h = \int_D f v_h \quad \forall v_h \in V_h. \tag{63}$$

Proposition 3.3. *Let u be the solution of problem (53) with F given by (62), and let $u_{0,h}$ be the solution of (63). There exists a constant $C>0$ depending only on the constants in (15) and (16) such that*

$$\mathbb{E}\left[\|\nabla(u-u_{0,h})\|^2_{L^2(D)}\right]^{\frac{1}{2}} \leq C\left[\eta_1^2+\eta_2^2\right]^{\frac{1}{2}},$$

with

$$\eta_1^2 := \sum_{K\in\mathcal{T}_h} h_K^2 \int_K (f+\Delta u_{0,h} - a_0 u_{0,h}^3)^2 + \sum_{e\in\mathcal{T}_h} h_e \int_e [\nabla u_{0,h}\cdot\mathbf{n_e}]^2_{\mathbf{n_e}}$$

$$\eta_2^2 := \varepsilon^2\sigma^2 \int_D \sum_{j=1}^{L} a_j^2 u_{0,h}^6.$$

Proof. The proof is based on the relations

$$\int_D \nabla(u-u_{0,h})\cdot\nabla v = \int_D fv - \int_D a_0 u_{0,h}^3 v - \int_D \nabla u_{0,h}\cdot\nabla v - \int_D (au^3 - a_0 u_{0,h}^3)v$$

and

$$-\int_D (au^3 - a_0 u_{0,h}^3)v = -\int_D a\int_0^1 3(u_{0,h}+t(u-u_{0,h}))^2(u-u_{0,h})dtv - \int_D (a-a_0)u_{0,h}^3 v.$$

Since a is positive, the first term of the right-hand side of the last equality is less or equal to zero for $v = u - u_{0,h}$. □

4 Comparison with the Stochastic Collocation method

We perform here a comparison of the computational costs between the SC-FEM method and the one presented here, called *perturbation method* in the sequel, when comparable accuracy is reached.

Shortly, the SC-FEM applied to the model problem (2) consists, given a set of (collocation) points $\{\mathbf{y}_k \in \Gamma, k = 1, ..., N_c\}$, in finding $u_h(\cdot, \mathbf{y}_k) \in V_h$ such that

$$\int_D a(\mathbf{x}, \mathbf{y}_k)\nabla u_h(\mathbf{x}, \mathbf{y}_k) \cdot \nabla v_h(\mathbf{x})d\mathbf{x} = \int_D f(\mathbf{x})v_h(\mathbf{x})d\mathbf{x} \quad \forall v_h \in V_h$$

for $k = 1, ..., N_c$ and build a global polynomial approximation

$$u_h^{SC}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N_c} u_h(\mathbf{x}, \mathbf{y}_k)\psi_k(\mathbf{y}),$$

for appropriate multivariate polynomials $\{\psi_k\}_{k=1}^{N_c}$. Since the FEM is used for the physical space approximation in both methods (stochastic collocation and *perturbation*), we use the same mesh for the discretization of D. For a comparable statistical error, say an error with convergence rate of order ε^2, we take $N = 1$ in the expansion (9) of u for the *perturbation method* and use a sparse grid of level 1 for the SC method, based either on Clenshaw-Curtis (see [37]) or Gaussian abscissas. The construction of the sparse grid interpolant of level 1 is briefly described in the following. We refer to [13, 15, 38] for more details and the general construction of sparse grid of arbitrary level. First, the sparse grid interpolant of level 0 of a function $f(\mathbf{y})$, denoted $S_0 f$, is simply the evaluation of the function at $(Y_1^0, ..., Y_L^0)$, where Y_j^0 is the unique interpolation point in direction j. Next, for each random variable $Y_j \in \Gamma_j$, we define the sequence of interpolation points at level $i \geq 1$ by $\{Y_{j,k}^i,\ k = 1, ..., m(i)\}$, where the number of collocation points $m(i)$ can be taken for instance as

$$m(i) = i + 1 \quad \text{or} \quad m(i) = \begin{cases} 1 & \text{if } i = 0 \\ 2^i + 1 & \text{if } i \geq 1. \end{cases}$$

The former choice for m corresponds to a total degree (TD) approximation space while the latter corresponds to a Smolyak one (see [9]). Notice that compared to the articles mentioned above, the level index i starts here at 0 instead of 1. We define then the one dimensional (Lagrange) interpolation operator in direction j at level $i = 1$ by

$$\mathcal{U}_j^1 f(Y_1, ..., Y_L) := \sum_{k=1}^{m(1)} f(Y_1^0, ..., Y_{j-1}^0, Y_{j,k}^1, Y_{j+1}^0, ..., Y_L^0) \left(\prod_{l=1, l\neq k}^{m(1)} \frac{Y_j - Y_{j,l}^1}{Y_{j,k}^1 - Y_{j,l}^1} \right),$$

which is a polynomial of degree $m(1) - 1$ in the random variable Y_j and constant in all other variables. Finally, the level 1 sparse grid interpolant is defined as

$$S_1 f := S_0 f + \sum_{j=1}^{L} (\mathcal{U}_j^1 f - S_0 f) = (1 - L)S_0 f + \sum_{j=1}^{L} \mathcal{U}_j^1 f$$

which is nothing else than the sum of the level 0 sparse grid interpolant and the details in each direction. The type of points in each direction is chosen according to the distribution of the random variables. Note that the use of Clenshaw-Curtis points, which are the extrema of Chebyshev polynomials and which are suitable for uniformly distributed random variables, and Smolyak sparse grid leads to nested set of abscissas. However, since only sparse grids of level 1 are considered, there is no real advantage to consider hierarchical sparse grids. In both cases $m(1) = 2$ and Gauss-Legendre abscissas and $m(1) = 3$ and Clenshaw-Curtis abscissas, referred to as SC1 and SC2 in the following, the sparse grid of level 1 consists of $2L + 1$ collocation points.

Let W_l, respectively W_{nl}, denotes the work to solve once a given linear, respectively nonlinear, problem. Moreover, let $W_{\tilde{l}}$ denotes the work to solve the linear problem for U_j associated to the nonlinear one, see (55). Table 1 contains the computational costs for the SC-FEM and the *perturbation method*. Notice that the work to construct the sparse grid is not taken into account.

	linear problem	nonlinear problem
SC-FEM	$(2L+1)\cdot W_l$	$(2L+1)\cdot W_{nl}$
perturbation method	$(L+1)\cdot W_l$	$W_{nl}+L\cdot W_{\tilde{l}}$

Table 1: Computational costs for the SC-FEM and the *perturbation method*.

The *perturbation method* presents no real advantage for solving linear problems since the costs for both methods differ only by a factor 2. The situation is different when a nonlinear problem is considered. Indeed, when using the SC method, we need to solve as many nonlinear problems as collocation points, i.e. $2L+1$ problems, whereas only one nonlinear problem needs to be solved for the *perturbation method*. The L remaining problems, to compute the U_j, $j=1,...,L$, are linear and so usually much cheaper to solve. However, one should invest extra effort to derive by hand the Fréchet derivatives and implement the problems solved by the U_j, $j=1,...,L$.

5 Numerical results

We now give some numerical examples in $1D$ to illustrate the theoretical estimators derived in the previous sections. Let $D=(0,1)$. The errors in $L^2_\rho(\Gamma;H^1_0(D))$ and $L^2_\rho(\Gamma;L^2(D))$ norms have been approximated with the standard Monte Carlo method, with a sample of size $K=10000$, i.e. for $V=H^1_0(D)$ or $L^2(D)$ we approximate

$$\|v(\mathbf{x},\mathbf{y})\|_{L^2_\rho(\Gamma;V)} \approx \left(\frac{1}{K}\sum_{k=1}^{K}\|v(\mathbf{x},\mathbf{y}_k)\|^2_V\right)^{\frac{1}{2}} \quad \forall v\in L^2_\rho(\Gamma;V),$$

where $\{\mathbf{y}_k\}\in\Gamma$ are i.i.d realizations of the random vector $\mathbf{y}$. With this choice for the sample size, the variance of the estimation of the error for all the considered values of h and ε is at most 10^{-5} the estimated error. Since the exact random solution of the problems considered below is not known, the error is computed with respect to a reference solution computed on a fine uniform mesh for D, namely with a mesh-grid of length $h_{ref}=2^{-12}$. Notice that if we take a FE space of mesh size $h=h_{ref}$, then only the statistical error is considered.

5.1 Linear problems

Let us first consider $L=50$ random variables Y_j, $j=1,...,L$, which can take the values ± 1 with probability $\frac{1}{2}$. Such discrete random variables have zero mean, unit variance and unit fourth moment. Similarly to what is done in [15], we take a diffusion coefficient of the form

$$a(x,\mathbf{y}) = 1+\varepsilon\sum_{j=1}^{L}\frac{\cos(2\pi jx)}{(\pi j)^2}Y_j(\omega),$$

which is similar to a (truncated) Karhunen-Loève expansion with eigenvalues of order $\frac{1}{j^4}$. With this choice of stochastic diffusion coefficient, we have $a_{min}=1-\frac{\varepsilon}{6}$ and $a_{max}=1+\frac{\varepsilon}{6}$. Finally, we consider two different right-hand sides, namely

$$f_1(x)=1 \quad \text{and} \quad f_2(x)=72\left(1-72(x-0.5)^2\right)e^{-36(x-0.5)^2}.$$

The latter corresponds to the exact solution $u_0(x)=e^{-36(x-0.5)^2}$ for Problem (10).

We show in Figure 1 the convergence rate of the error $u-u_{0,h}$ in the $L^2_\rho(\Gamma;H^1_0(D))$-norm, along with the *a posteriori* estimator given by (33), with respect to $2^{-9}\le h\le 2^{-3}$ for $\varepsilon=32h$. Based on this result, we can see that a division of h and ε by two halves the error, which is in agreement with the convergence of $\|u-u_{0,h}\|_{L^2_\rho(\Gamma;H^1_0(D))}$ in $\mathcal{O}(h+\varepsilon)$ predicted by the estimator (33).

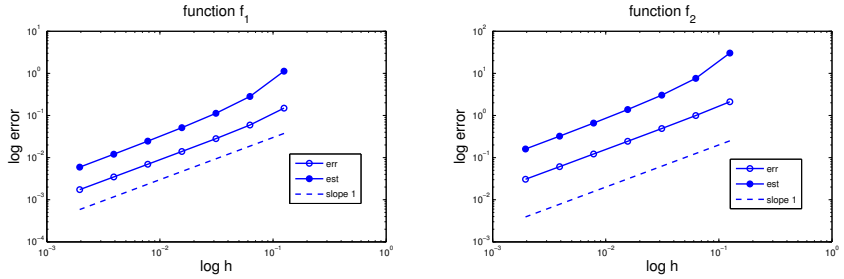

Figure 1: Convergence orders for problem (2) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and $u_{0,h}$ in $L^2_\rho(\Gamma; H^1_0(D))$-norm w.r.t h with $\varepsilon = 32h$.

If we consider now the error in L^2-norm in space, we should get a convergence of order h^2 for $\varepsilon = Ch^2$. Figure 2, which contains the plot of the error and estimator (36) for $C = 2048$ and $2^{-9} \leq h \leq 2^{-5}$, confirms that this is also the case.

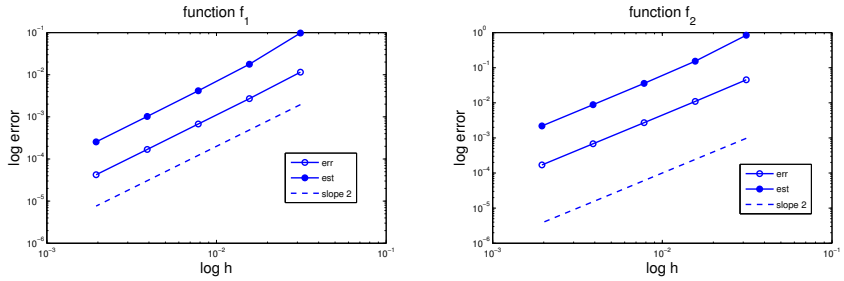

Figure 2: Convergence orders for problem (2) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and $u_{0,h}$ in $L^2_\rho(\Gamma; L^2(D))$-norm w.r.t h with ε fixed to $2048h^2$.

Concerning the convergence rate of the second order approximation, we present on Figure 3 the error between u and u^1_h in $L^2_\rho(\Gamma; H^1_0(D))$-norm with respect to $2^{-5} \leq \varepsilon \leq 2^{-1}$ for $h = \varepsilon^2$. This

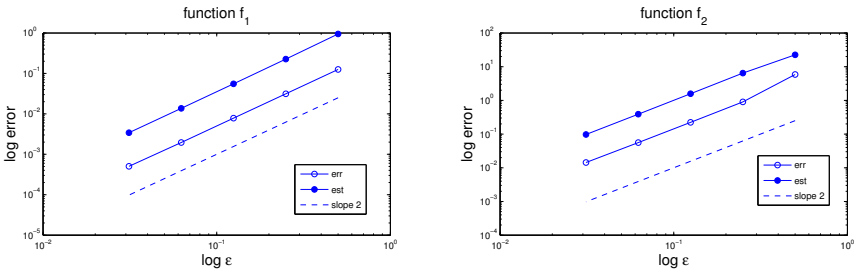

Figure 3: Convergence orders for problem (2) with $f = f_1$ (left) and $f = f_2$ (right). Log log scale plot of the error between u and u^1_h in $L^2_\rho(\Gamma; H^1_0(D))$-norm w.r.t ε with $h = \varepsilon^2$.

result confirms the convergence in $\mathcal{O}(\varepsilon^2)$ of the stochastic truncation predicted by (44), when the exact solution is approximated by $u_0 + \varepsilon u_1$.

Finally, the convergence rate of the error for the first and second order approximation with respect to h in the $L^2_\rho(\Gamma; H^1_0(D))$-norm for several given (fixed) values of ε is depicted in Figure 4. First, we can notice that a better accuracy is reached when u is approximated by u^1_h than

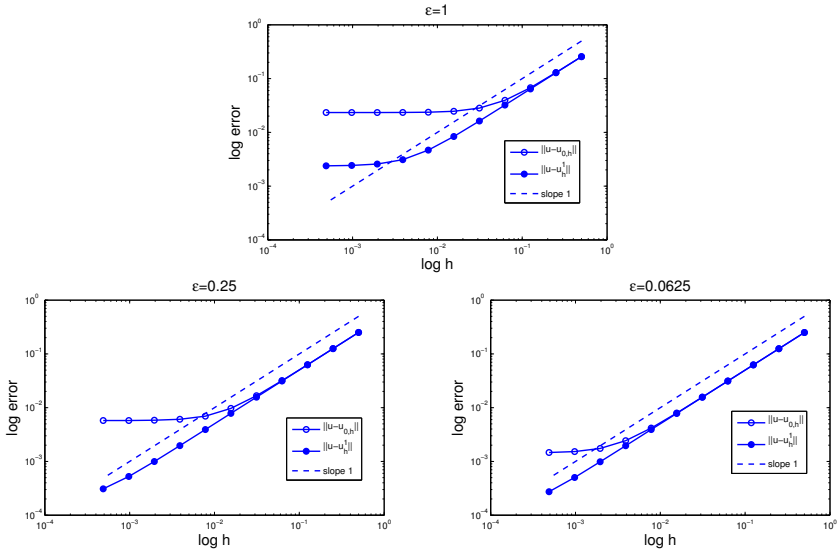

Figure 4: Convergence rate for problem (2) with $f = f_1$ for $\varepsilon = 1$ (top), $\varepsilon = 2^{-2}$ (bottom left) and $\varepsilon = 2^{-4}$ (bottom right). Log log scale plot of the error in $L^2_\rho(\Gamma; H^1_0(D))$-norm w.r.t h.

with only the deterministic part $u_{0,h}$, except for coarse meshes where the FE error is dominating yielding comparable accuracy. Moreover, the global approximation error remains constant for mesh sizes smaller than a critical value h_0 of the mesh-size. Any further refinement of the mesh below this value should thus be avoided since it would not improve the global approximation error, being dominated by the stochastic error.

Based on this observation, it is interesting to determine how fine the mesh should be to get a comparable error in h and ε. More precisely, for a given ε, we would like to find h such that

$$\frac{T-1}{T}\eta_2 \leq \eta_1 \leq \frac{T+1}{T}\eta_2 \tag{64}$$

for a given $T > 1$. This can be done in 1D using Algorithm 1 given below, where $N_h + 1$ denotes the number of discretization points in $[0, 1]$. Algorithm 1 only uses uniform refinements/coarsening. Of course, adaptive refinements could be considered as well exploiting the local nature of the estimator η_1 in (28).

Applying Algorithm 1 to our problem for $T = 2$ and various given ε, we get the results presented in Table 2.

	f_1			f_2		
ε	N	η_1	η_2	N	η_1	η_2
1	64	0.03125	0.02299	256	0.22264	0.19669
0.5	128	0.01563	0.01150	512	0.11132	0.09834
0.25	256	0.00781	0.00575	1024	0.05566	0.04917
0.125	512	0.00391	0.00288	2048	0.02783	0.02458
0.0625	1024	0.00195	0.00144	4096	0.01392	0.01229

Table 2: Value of $h = N_h^{-1}$ with respect to ε such that (64) holds with $T = 2$.

Remark 5.1. *Similar results are obtained when independent uniformly distributed random variables in* $[-\sqrt{3}, \sqrt{3}]$ *are considered. Notice that in this case, the random variables still have zero mean*

Algorithm 1 find $h = N_h^{-1}$ such that (64) holds

Require: N_{init} and T
Ensure: mesh-size h which yield comparable accuracy in h and ε
1: $N_h = N_{init}$
2: Compute $u_{0,h}$ on the uniform partition $x_i = ih$, $h = N_h^{-1}$, $i = 0, 1, ..., N_h$
3: Compute η_1 and η_2 according to (28)
4: **if** $\frac{T-1}{T} \leq \frac{\eta_1}{\eta_2} \leq \frac{T+1}{T}$ **then**
5: stop
6: **else**
7: **if** $\frac{\eta_1}{\eta_2} < \frac{T-1}{T}$ **then**
8: $N_h \leftarrow \lfloor \frac{N_h}{2} \rfloor$ (mesh too fine)
9: **else**
10: $N_h \leftarrow 2N_h$ (mesh too coarse)
11: **end if**
12: go to 2.
13: **end if**

and unit variance but $\mathbb{E}[Y_j^4] = \frac{9}{5}$. *This only modifies the part* η_3 *in the a posteriori error estimator (44) for* $\|u - u_h^1\|_{L^2_\rho(\Gamma;H_0^1(D))}$. *Moreover, the lower and upper bound for the diffusion coefficient is given respectively by* $a_{min} = 1 - \frac{\sqrt{3}\varepsilon}{6}$ *and* $a_{max} = 1 + \frac{\sqrt{3}\varepsilon}{6}$ *in this case.*

5.2 Comparison with Stochastic Collocation method

We finally illustrate the findings of Section 4 concerning the computation costs for the SC-FEM and the *perturbation method*. We consider the linear problem (2) and the nonlinear problem (53) with F given by (56). In both cases, homogeneous Dirichlet boundary condition are considered and we assume that the random variables Y_j, $j = 1, ..., L$, that appear in the characterization of a (3) are uniform random variables in $[-\sqrt{3}, \sqrt{3}]$. We compare the computation time to solve the two problems with accuracy of order 2 in ε. Such accuracy is reached when we consider a sparse grid of level 1 for the SC-FEM method and the second order approximation $u \approx u_{0,h} + \varepsilon u_{1,h}$ for the *perturbation method*. Note that $u_{1,h} = \sum_{j=1}^{L} U_{j,h} Y_j$ where $U_{j,h}$ for $j = 1, ..., L$ is the solution of

$$\int_D a_0 \nabla U_{j,h} \cdot \nabla v_h + \int_D 3u_{0,h}^2 U_{j,h} v_h = -\int_D a_j \nabla u_{0,h} \cdot v_h \quad \forall v_h \in V_h.$$

when problem (53) is considered. Finally, we use the same physical space discretization for both methods, namely a uniform partition with $h = 2^{-12}$. With this choice of mesh size, the work to solve the $(2L + 1)$ problems dominates the one needed to construct the grid. The computational time to solve both problems with respect to the number of random variables L is given on Figure 5.

As predicted in section 4, the *perturbation method* presents no real advantage in terms of computation time over the Stochastic Collocation one, since it is only twice faster. This factor 2 comes from the fact that the *perturbation method* requires the solution of $L + 1$ problems, while $2L + 1$ problems need to be solve in the Stochastic Collocation method. The situation is different for nonlinear problems. In this case, the *perturbation method* is significantly faster than the Stochastic Collocation one. Indeed, only one nonlinear problem and L linear problems need to be solve for the former, to obtain respectively the deterministic part u_0 of u and the U_j, $j = 1, ..., L$. For the SC method, we need to solve as many nonlinear problems as collocation points. Even for the nonlinear problem considered here, where the nonlinearity comes from the term u^3 and which is quite cheap to solve, the *perturbation method* is about 8 times faster.

To conclude, we can mention that for $h = h_{ref}$, i.e. without error due to FE approximation and a convergence of the error in $\mathcal{O}(\varepsilon^2)$, the error for the *perturbation method* is about 1.4 and

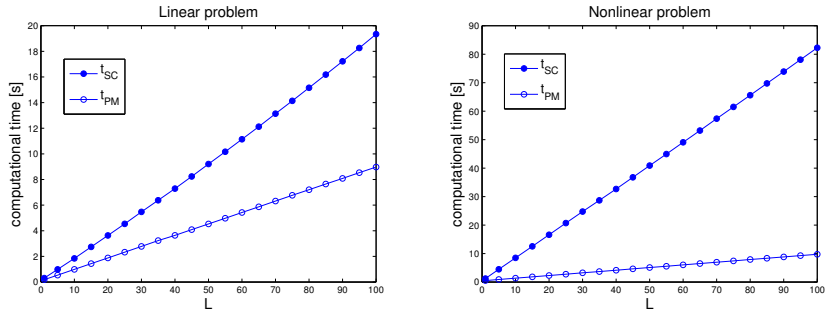

Figure 5: Time to solve the linear problem (2) and the nonlinear problem (56) with accuracy of order 2 in ε using the SC-FEM and the *perturbation method*.

3.5 times larger than the error obtained using respectively SC1 and SC2. However, for a given problem, that is for fixed value of ε and L, the CPU time with respect to the error is lower for the *perturbation method*, as shown on Figure 6 for problems (2) and (56) with $f = f_2$, $\varepsilon = 0.5$, $L = 10$ and $h^{-11} \leq h \leq h^{-3}$. Notice that the results for SC1 are not depicted on this figure since they are indistinguishable from those of SC2.

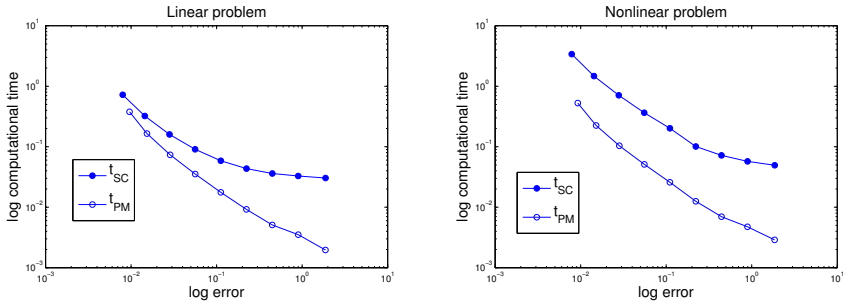

Figure 6: Log log scale plot of the computational time w.r.t. the error in $L^2_\rho(\Gamma; H^1_0(D))$-norm using the SC-FEM with Smolyak and Clenshaw-Curtis abscissas and the *perturbation method*.

6 Conclusions

In this paper, we have performed error analyses for elliptic PDEs with coefficients affected by small uncertainties, characterized through random variables. The exact random solution has been approximated using a perturbation approach combined with the finite element method for the physical space discretization.

For the first order approximation $u \approx u_{0,h}$, we derived strong and weak *a priori* error estimates as well as *a posteriori* error estimates in the $L^2(\Omega; H^1_0(D))$ and $L^2(\Omega; L^2(D))$ norms. These estimators naturally split into two parts, namely the error in h due to the physical discretization and the error in ε due to the model. In the *a priori* error estimation, we have shown that the order of the weak error in the model is twice the order of the strong error, the order of the error due to FE approximation being the same in both cases. The *a posteriori* error estimator in $L^2(\Omega; H^1_0(D))$ norm that we have obtained is a computable quantity of order $h + \varepsilon$. Given $u_{0,h}$, this estimator is cheap to compute and does not require any other FE solution. It can be used for mesh adaptation so that comparable accuracy in h and ε is reached. We have shown that taking the L^2 norm in physical space leads to a gain of one order in h but no improvement in the error due to the model.

Finally, we gave the sketch of the derivation of a goal-oriented estimator, which is more suitable than an estimator in global norm when a particular quantity of interest is considered.

The *a posteriori* error estimation procedure for the error in the $L^2(\Omega; H_0^1(D))$ norm has been applied to the second-order approximation $u \approx u_{0,h} + \varepsilon u_{1,h}$, before giving a generalization for approximations of any order.

A posteriori error estimates have then been derived for a class of nonlinear problems through three different examples. A comparison in terms of computational costs with the Stochastic Collocation method has been performed, considering an error of order 2 in the model. The *perturbation method* presents only mild advantages for solving linear problems, the computational cost being halved with respect to the SC method. The situation is different for nonlinear problems. Indeed, the SC method requires the resolution of as many nonlinear problems as collocation points while for the *perturbation method*, only one nonlinear problem has to be solved for $u_{0,h}$, the remaining problems being linear.

References

[1] G.S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications.* Springer Series in Operations Research and Financial Engineering, Springer-Verlag, New-York, 1996.

[2] R.E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica, Cambridge University Press, pp. 1-49*, 1998.

[3] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, and I. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *Journal of Computational Physics*, 230(10):3668–3694, 2011.

[4] A. Barth, C. Schwab, and N. Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161, September 2011.

[5] K.A. Cliffe, M.B. Giles, R. Scheichl, and A.L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput Visual Sci*, (14):3–15, 2011.

[6] S. Heinrich. *Multilevel Monte Carlo Methods.* Lecture notes in Comput. Sci. 2179, Springer-Verlag, Berlin, pp. 3624-3651, 2001.

[7] I. Babuska, 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.

[8] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. Convergence of quasi-optimal Stochastic Galerkin methods for a class of PDEs with random coefficients. *Comput. Math. Appl.*, 67(4):732–751, 2014.

[9] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. Stochastics spectral Galerkin and collocation methods for PDEs with random coefficients: A numerical comparison. *Lecture Notes in Computational Science and Engineering, Springer*, 76:43–62, 2011.

[10] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach.* Springer, New York, 1991.

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

[12] I. Babuska, 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.

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

[14] 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. Numer. Anal.*, 46(5):2411–2442, 2008.

[15] 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.

[16] M. Eigel, C. Gittelson, C. Schwab, and E. Zander. Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mehc. Engrg.*, (270):247–269, 2014.

[17] M. Eigel, C. Gittelson, C. Schwab, and E. Zander. A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes. *Tech. Report 2014-01, Seminar for Applied Mathematics, ETH Zürich*, 2014.

[18] L. Mathelin and O. Le Maître. Dual-based a posteriori error estimate for stochatic finite element methods. *Commun. Appl. Math. and Comp. Sci.*, 2(1):83–115, 2007.

[19] T. Butler, C. Dawson, and T. Wildey. A posteriori error analysis of stochastic spectral methods. *SIAM J. Sci. Comput.*, 33(3):1267–1291, 2011.

[20] C.M. Bryant, S. Prudhomme, and T. Wildey. A posteriori error control for partial differential equations with random data. *ICES REPORT 13-08, The Institute for Computational Engineering and Sciences, The University of Texas at Austin*, April 2013.

[21] R.C. Almeida and J.T. Oden. Solution verification, goal-oriented adaptive methods for stochastic advection-diffusion problems. *Comput. Mthods Appl. Mech. Engrg.*, 199:2472–2486, 2010.

[22] M. Kleiber and T.D. Hien. *The Stochastic Finite Element Method: Basic Perturbation Technique and Computer Implementation.* Wiley, Chichester, 1992.

[23] D.G. Cacuci. *Sensitivity and uncertainty analysis: Theory, Vol. 1.* Chapman & Hall/CRC, Boca Raton, 2003.

[24] A. Bonito, R.A. Devore, and R.H. Nochetto. Adaptive finite element methods for elliptic problems with discontinuous coefficients. *SIAM J. Numer. Anal.*, 51(6):3106–3134, 2013.

[25] I. Babuska and P. Chatzipantelidis. On solving elliptic stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 191(37-38), August 2002.

[26] C. Bernardi and R. Verfürth. Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numer. Math.*, 85:579–608, 2000.

[27] M. Loève. *Probability Theory I.* Spinger-Verlag, New-York, 4th ed., Gaduate Texts in Mathematics, Vol. 45, 1977.

[28] M. Loève. *Probability Theory II.* Spinger-Verlag, New-York, 4th ed., Gaduate Texts in Mathematics, Vol. 46, 1978.

[29] J. Beck, F. Nobile, L. Tamellini, and R. Tempone. On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods. *Math. Models Methods Appl. Sci.*, 22(9):1250023–1–1250023–33, 2012.

[30] P.G. Ciarlet. *The Finite Element method for elliptic problems.* North Holland, Amsterdam, 1978.

[31] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New-York, 3th ed., Texts in Applied Mathematics, Vol. 15, 2008.

[32] P. Clément. Approximation by finite element functions using local regularization. *RAIRO*, (R-2):77–84, August 1975.

[33] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, 1996.

[34] W. Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Birkhäuser Verlag, Basel, 2003.

[35] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

[36] J.T. Oden and S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. *Computers and Mathematics with Applications*, 41:735–756, 2001.

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

[38] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18:209–232, 1998.

Recent publications:

MATHEMATICS INSTITUTE OF COMPUTATIONAL SCIENCE AND ENGINEERING
Section of Mathematics
Ecole Polytechnique Fédérale
CH-1015 Lausanne

28.2014 LAURA IAPICHINO, ALFIO QUARTERONI, GIANLUIGI ROZZA, STEFAN VOLKWEIN:
Reduced basis method for the stokes equations in decomposable parametrized domains using greedy optimization

29.2014 ASSYR ABDULLE, PATRICK HENNING:
Localized orthogonal decomposition method for the wave equation with a continuum of scales

30.2014 DANIEL KRESSNER, ANDRÉ USCHMAJEW:
On low-rank approximability of solutions to high-dimensional operator equations and eigenvalue problems

31.2014 ASSYR ABDULLE, MARTIN HUBER:
Finite element heterogeneous multiscale method for nonlinear monotone parabolic homogenization problems

32.2014 ASSYR ABDULLE, MARTIN HUBER, GILLES VILMART:
Linearized numerical homogenization method for nonlinear monotone parabolic multiscale problems

33.2014 MARCO DISCACCIATI, PAOLA GERVASIO, ALFIO QUARTERONI:
Interface control domain decomposition (ICDD) methods for heterogeneous problems

34.2014 ANDRE USCHMAJEW:
A new convergence proof for the high-order power method and generalizations

35.2014 ASSYR ABDULLE, ONDREJ BUDÁC:
A Petrov-Galerkin reduced basis approximation of the Stokes equation in parametrized geometries

36.2014 ASSYR ABDULLE, MARTIN E. HUBER:
Error estimates for finite element approximations of nonlinear monotone elliptic problems with application to numerical homogenization

37.2014 LARS KARLSSON, DANIEL KRESSNER, ANDRÉ USCHMAJEW:
Parallel algorithms for tensor completion in the CP format

38.2014 PAOLO TRICERRI, LUCA DEDÈ, SIMONE DEPARIS, ALFIO QUARTERONI, ANNE M. ROBERTSON, ADÉLIA SEQUEIRA:
Fluid-structure interaction siumlations of cerebral arteries modeled by isotropic and anistropic constitutive laws

39.2014 FRANCESCA BONIZZONI, FABIO NOBILE, DANIEL KRESSNER:
Tensor train approximation of moment equations for the log-normal Darcy problem

40.2014 DIANE GUIGNARD, FABIO NOBILE, MARCO PICASSO:
A posteriori error estimations for elliptic partial differential equations with small uncertainties