Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs [*]

Abdellah Chkifa, Albert Cohen, Ronald DeVore and Christoph Schwab

June 17, 2011

Abstract

The numerical approximation of parametric partial differential equations is a computational challenge, in particular when the number of involved parameter is large. This paper considers a model class of second order, linear, parametric, elliptic PDEs on a bounded domain D with diffusion coefficients depending on the parameters in an affine manner. For such models, it was shown in [11, 12] that under very weak assumptions on the diffusion coefficients, the entire family of solutions to such equations can be simultaneously approximated in the Hilbert space $V = H_0^1(D)$ by multivariate sparse polynomials in the parameter vector y with a controlled number N of terms. The convergence rate in terms of N does not depend on the number of parameters in V, which may be arbitrarily large or countably infinite, thereby breaking the curse of dimensionality. However, these approximation results do not describe the concrete construction of these polynomial expansions, and should therefore rather be viewed as benchmark for the convergence analysis of numerical methods. The present paper presents an adaptive numerical algorithm for constructing a sequence of sparse polynomials that is proved to converge toward the solution with the optimal benchmark rate. Numerical experiments are presented in large parameter dimension, which confirm the effectiveness of the adaptive approach.

1 Introduction

We consider parametric partial differential equations of the general form

$$\mathcal{D}(u, y) = 0$$

where $u \mapsto \mathcal{D}(u, y)$ is a partial differential operator that depends on a vector of parameters y, and therefore so does the solution $u = u(y)$. Parametric problems of this type arise in modeling complex systems in various contexts:

[*] This research was supported by the Office of Naval Research Contracts ONR-N00014-08-1-1113, ONR N00014-09-1-0107, the AFOSR Contract FA95500910500, the ARO/DoD Contracts W911NF-05-1-0227 and W911NF-07-1-0185, the National Science Foundation Grant DMS 0915231; the excellency chair of the Foundation "Science Mathématiques de Paris" awarded to Ronald DeVore in 2009. This publication is based on work supported by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). This research is also supported by the Swiss National Science Foundation under Grant SNF 200021-120290/1 and by the European Research Council under grant ERC AdG247277. CS acknowledges hospitality by the Hausdorff Institute for Mathematics, Bonn, Germany.

- Stochastic modelling: the parameters y are realizations of random variables which reflects the fact that the diffusion coefficient is not known exactly and is therefore modelled as a random field. The user is interested in the resulting statistical properties of the solution u. This is the point of view adopted for example in [16, 21, 2, 15, 1, 24, 25]. Here, the parameter y is generally infinite dimensional and any numerically viable approximation must necessarily address the issue of dimensional reduction, in addition to that of discretization.

- Deterministic modelling: the parameters y are known or controlled by the user, who is interested in studying the dependence of u with respect to these parameters for various purposes (for example, optimizing an output of the equation with respect to y). This is the point of view adopted for example in [7, 22].

In both of these settings, the main computational challenge is to *simultaneously* solve the entire parametric family of equations up to a prescribed accuracy ε, with reasonable computation cost. This task is particularly difficult when the number of involved parameters is large or if there are countably many parameters, i.e. when the dimension of the parameter domain is inifinite. In this paper, we place ourselves in this last situation.

More precisely, we consider the model parametric elliptic boundary value problem

$$-\mathrm{div}(a\nabla u) = f \ \text{ in } \ D \subset \mathbb{R}^d, \ \ u = 0 \ \text{ on } \ \partial D, \tag{1.1}$$

where $f \in H^{-1}(D)$ and $a(x,y) := \bar{a}(x) + \sum_{j\geq 1} y_j \psi_j(x)$, for $y = (y_j)_{j\geq 1} \in U := [-1,1]^{\mathbb{N}}$, where $\bar{a}$ and $(\psi_j)_{j\geq 1}$ are functions in $L^\infty(D)$. In the stochastic context, the series $\bar{a} + \sum_{j\geq 1} y_j \psi_j$ could result, for example, from a Karhúnen-Loève expansion of a random field $a(x,\omega)$, see [27].

We assume at a minimum that the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$ is bounded. By rearranging indices, if necessary, we may assume without loss of generality that $\|\psi_j\|_{L^\infty(D)}$, $j = 1, 2, \ldots$, is a non-increasing sequence. We work under the (minimal) *uniform ellipticity assumption* $\mathbf{UEA}(r, R)$ for some $0 < r \leq R < \infty$:

$$0 < r \leq \bar{a}(x) + \sum_{j\geq 1} y_j \psi_j(x) \leq R, \ \ x \in D, \, y \in U. \tag{1.2}$$

Observe that the lower inequality in this assumption is equivalent to

$$\sum_{j\geq 1} |\psi_j(x)| \leq \bar{a}(x) - r, \quad x \in D. \tag{1.3}$$

Assumption $\mathbf{UEA}(r, R)$ ensures existence and uniqueness of the solution $u(y)$ in $V = H_0^1(D)$, for all $y \in U$, with the a-priori estimate

$$\|u(y)\|_V \leq C_0 := \frac{\|f\|_{V^*}}{r},$$

where $\|v\|_V := \|\nabla v\|_{L^2(D)}$. We also introduce the average energy norm

$$\|v\|_{\bar{a}} := \Big(\int\limits_D \bar{a}|\nabla v|^2\Big)^{1/2},$$

which, under $\mathbf{UEA}(r, R)$, is equivalent to the $H_0^1(D)$ norm $\|\cdot\|_V$ in the sense that

$$\sqrt{r}\|v\|_V \leq \|v\|_{\bar{a}} \leq \sqrt{R}\|v\|_V, \ \ v \in V.$$

We are interested in approximating the map $y \mapsto u(y)$ by multivariate polynomials in y with coefficients in V. One way to obtain such an approximation is by truncating in some way the Taylor expansion of u at

$y = 0$. As in [11, 12], we denote by $\mathcal{F}$ the (countable) set of all sequences of nonnegative integers which are *finitely supported* (i.e. those sequence for which only finitely many terms are nonzero), we are thus interested in the summability properties in V of partial sums of the formal Taylor series $\sum_{\nu\in\mathcal{F}} t_\nu y^\nu$, where for each $\nu = (\nu_j)_{j\geq 1} \in \mathcal{F}$, we define

$$y^\nu := \prod_{j\geq 1} y_j^{\nu_j} \ \text{ and } \ t_\nu := \frac{1}{\nu!}\partial^\nu u(0) \in V \ \text{ with } \ \nu! := \prod_{j\geq 1} \nu_j! \ \text{ and } \ 0! := 1.$$

In [12] we proved the following result (Theorem 1.2 in [12]).

Theorem 1.1 *If $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$ and if* $\mathbf{UEA}(r, R)$ *holds, then*

$$u(y) = \sum_{\nu\in\mathcal{F}} t_\nu y^\nu, \tag{1.4}$$

in the sense of unconditional convergence in $L^\infty(U, V)$ and $(\|t_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$.

Analogous results for other types of partial differential equations are available in [19, 20]. Theorems such as 1.1 have important implications concerning the efficient numerical approximation of the parametric solution map $y \mapsto u(y)$ which, in turn, opens a perspective for novel computational approaches in parameter identification problems [26]. To describe these, for any sequence $(a_\nu)_{\nu\in\mathcal{F}}$ of real numbers and any $k \geq 1$, we define the sets $\Lambda_k^* := \Lambda_k^*((a_\nu)_{\nu\in\mathcal{F}})$ of k largest elements in absolute value. The sets Λ_k^* are generally not unique because of possible ties in the size of the $|a_\nu|$. However, if $(a_\nu)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$ and if Λ_k^* is any of these sets, then for any $q > p$

$$(\sum_{\nu\notin\Lambda_k^*} |a_\nu|^q)^{1/q} \leq \|(a_\nu)\|_{\ell_p(\mathcal{F})} k^{-\frac{1}{p}+\frac{1}{q}}. \tag{1.5}$$

We refer either to [13] or §3.3 in [12] for a proof of this elementary fact.

Working under the assumptions of the above theorem, for each $k = 1, 2, \ldots$, we denote by $\Lambda_k^* \subset \mathcal{F}$ the set of indices $\nu \in \mathcal{F}$ corresponding to the k largest of the $\|t_\nu\|_V$, with ties broken in an arbitrary (but consistent) way. We then have

$$\sup_{y\in U} \|u(y) - \sum_{\nu\in\Lambda_k^*} t_\nu y^\nu\|_V \leq \sum_{\nu\notin\Lambda_k^*} \|t_\nu\|_V \leq \|(\|t_\nu\|_V)\|_{\ell_p(\mathcal{F})} k^{-s}, \quad s := \frac{1}{p} - 1 \tag{1.6}$$

The sparse polynomials $\sum_{\nu\in\Lambda_k^*} t_\nu y^\nu$ therefore provide a simultaneous approximation of the family $\{u(y)\ ;\ y \in U\}$ at the cost of computing k functions $t_\nu \in V$. Quite remarkably the rate k^{-s} and the constant in (1.6) is independent of the number of parameters y_j which may be countably infinite. Thus, (1.6) implies that one can in principle overcome the curse of dimensionality in the approximation of $u(y)$.

In computation, however, the sets Λ_k^* in (1.6) are not known to us and to find them we would ostensibly have to compute all of the t_ν which is infeasible. To obtain computable sequences of index sets, we shall not insist on optimality: we shall say that *a nested sequence $(\Lambda_n)_{n\geq 0}$ of finite subsets $\Lambda_n \subset \mathcal{F}$ is near optimal* if there is a constant $C \geq 1$ such that for s as in (1.6) and for every $n \geq 0$

$$\sum_{\nu\notin\Lambda_n} \|t_\nu\|_V \leq C\|(\|t_\nu\|_V)\|_{\ell_p(\mathcal{F})} (\#(\Lambda_n))^{-s}. \tag{1.7}$$

The goal of the present paper is to give a concrete algorithm that adaptively builds a near optimal sequence $(\Lambda_n)_{n\geq 0}$ and corresponding Taylor coefficients $(t_\nu)_{\nu\in\Lambda_n}$ at a cost that scales linearly in $\#(\Lambda_n)$. We should point out that a similar program was developed when solving a *single* PDE by either adaptive wavelet methods [9, 10, 17] or by adaptive finite element methods [14, 23, 5, 28]. In these papers, it was proved that certain iterative refinement algorithms based on a-posteriori analysis generate adaptive wavelet sets or adaptive meshes such that the approximate solution converges with the optimal rate allowed by the exact solution. A common point between these algorithms and the one that we study in this paper is the use of a *bulk chasing procedure* in order to build the set Λ_{n+1} from the set Λ_n. However, our present setting is significantly different, since the index sets Λ_n are picked from the infinite dimensional lattice $\mathcal{F}$ and the coefficients associated to each $\nu \in \Lambda_n$ are functions in V instead of numbers.

Our paper is organized as follows. In §2, we show that the sets Λ_n may be picked from a restricted class called *monotone sets* while retaining the optimal rate.

We show in §3 how the Taylor coefficients associated to a monotone set may be recursively computed by solving one elliptic boundary value problem at a time per coefficient, and we establish a useful estimate for the energy of the Taylor coefficients outside a monotone set.

A first adaptive algorithm is proposed in §4 and proved to converge with the optimal rate in the sense that the sets Λ_n generated by the algorithm satisfy (1.7). A defect of this algorithm is that the bulk chasing procedure at step n requires the computation of the $\|t_\nu\|_{\bar{a}}$ for ν in a neighbourhood of Λ_n which has infinite cardinality and is therefore not practical.

We remedy this defect in §5 by introducing a second algorithm which operates the bulk search on a finite set, and which is also proved to converge with the optimal rate s in (1.6).

We study in §6 the additional error which is induced on the approximation of the map $y \mapsto u(y)$ by the spatial discretization when solving the boundary value problems that give the Taylor coefficients, for example by a finite element method on D. We prove that the additional error introduced by the finite element discretization of the coefficients is independent of the number of computed Taylor coefficients.

Finally, numerical experiments are presented in §7, for a finite but high dimensional test case ($y \in [-1,1]^{64}$), and using finite element for the spatial discretization. We test the adaptive bulk search strategy, and compare it with non-adaptive strategies based on a-priori choices of the sets Λ_n. These experiments confirm the superiority of the adaptive approach. We also propose alternate adaptive strategies which are computationally much cheaper than the bulk search and exhibit, in our numerical examples, the same convergence rate as the approximations obtained by bulk search yet without complete theoretical justification. In the practically relevant case where the goal of computation is to compute an average in y of the solution (corresponding to an expectation of the random solution) we show that the results based on our adaptive algorithm strongly outperform those using the Monte-Carlo method

2 Monotone sets

In this section, we give a finer description of the approximation properties of the Taylor series by introducing the notion of *monotonicity*. This notion is based on the following ordering of $\mathcal{F}$: for $\mu, \nu \in \mathcal{F}$, $\mu \le \nu$ if and only if $\mu_j \le \nu_j$ for all $j \ge 1$. We will also say that $\mu < \nu$ if and only if $\mu \le \nu$ and $\mu_j < \nu_j$ for at least one value of j.

Definition 2.1 *A sequence $(a_\nu)_{\nu \in \mathcal{F}}$ of nonnegative real numbers is said to be* monotone decreasing *if and only if for all $\mu, \nu \in \mathcal{F}$*

$$\mu \le \nu \Rightarrow a_\nu \le a_\mu \; .$$

A non empty set $\Lambda \subset \mathcal{F}$ is called monotone *if and only if $\nu \in \Lambda$* and *$\mu \le \nu \Rightarrow \mu \in \Lambda$. For a monotone set*

$\Lambda \subset \mathcal{F}$, we define its margin $\mathcal{M} = \mathcal{M}(\Lambda)$ as follows:

$$\mathcal{M}(\Lambda) := \{\nu \notin \Lambda \; ; \; \exists j > 0 \; : \nu - e_j \in \Lambda\} \,, \tag{2.1}$$

where $e_j \in \mathcal{F}$ is the Kronecker sequence: $(e_j)_i = \delta_{ij}$ for $i, j \in \mathbb{N}$.

Notice that the margin $\mathcal{M}(\Lambda)$ is an infinite set even when Λ is finite since there are infinitely many variables. Any nonempty monotone set contains the null index $(0, 0, \cdots)$, which we will denote in what follows with slight abuse of notation by 0. Intersections and unions of monotone sets are also monotone. Also, note that $\Lambda \cup \mathcal{M}(\Lambda)$ is a monotone set.

For any $\nu \in \mathcal{F}$, we let $|\nu| := \sum_{i \geq 1} \nu_i$. We say that ν is maximal in a set $\Lambda \subset \mathcal{F}$ if and only if there exists no $\mu > \nu$ in Λ. If $\Lambda \subset \mathcal{F}$ satisfies $N := N(\Lambda) := \max_{\nu \in \Lambda} |\nu| < \infty$, then any $\nu \in \Lambda$ for which $|\nu| = N$ is a maximal element. In particular, any finite set Λ has at least one maximal element. If Λ is monotone and if ν is maximal in Λ, then $\Lambda - \{\nu\}$ is monotone.

Remark 2.2 *If $(a_\nu)_{\nu \in \mathcal{F}}$ is a monotone sequence, the set $\Lambda_k^* = \Lambda_k^*((a_\nu)_{\nu \in \mathcal{F}})$ of indices corresponding to the k-largest a_ν in absolute value is always a monotone set whenever it is unique: it is then equivalently given by $\Lambda_k^* = \{\nu \in \mathcal{F} : a_\nu \geq \eta\}$ for some threshold η which depends on k. In the case of non-uniqueness, there exists at least one realization of a Λ_k^* which is monotone. We refer to such a set as a monotone realization of Λ_k^*. Such a realization may be constructed as follows: consider the largest threshold η such that the monotone set $\{\nu \in \mathcal{F} : \;\; a_\nu \geq \eta\}$ has more than k elements, and trim this set by removing iteratively a maximal ν until it has exactly k elements.*

Remark 2.3 *We localize the notion of monotone sequences and monotone sets as follows: if $\mathcal{F}_0 \subset \mathcal{F}$ is any subset, we say that the sequence $(a_\nu)_{\nu \in \mathcal{F}}$ is monotone on $\mathcal{F}_0$ (or that $(a_\nu)_{\nu \in \mathcal{F}_0}$ is monotone) if and only if*

$$\mu, \nu \in \mathcal{F}_0 \;\text{ and }\; \mu \leq \nu \Rightarrow a_\nu \leq a_\mu.$$

Clearly, a monotone sequence is monotone on any set $\mathcal{F}_0$. Likewise we say that a subset $\mathcal{F}_1 \subset \mathcal{F}_0$ is monotone in $\mathcal{F}_0$ if and only if

$$\nu \in \mathcal{F}_1, \; \mu \in \mathcal{F}_0 \;\text{ and }\; \mu \leq \nu \Rightarrow \mu \in \mathcal{F}_1.$$

In the case where $\mathcal{F}_0$ is monotone, this is equivalent to saying that $\mathcal{F}_1$ is monotone. If (a_ν) is monotone on $\mathcal{F}_0$, a set of indices corresponding to the k-largest a_ν in absolute value with $\nu \in \mathcal{F}_0$ is monotone in $\mathcal{F}_0$ whenever it is unique. If it is not unique, there exists at least one realization of such a set which is monotone. This set may be obtained by the same trimming procedure as in Remark 2.2.

The *monotone majorant of a bounded sequence* $(a_\nu)_{\nu \in \mathcal{F}}$ is the sequence

$$\mathbf{a}_\nu := \max_{\mu \geq \nu} |a_\mu|, \quad \nu \in \mathcal{F} \,.$$

We define $\ell^p_m(\mathcal{F})$ as the set of all sequences which have their monotone majorant in $\ell^p(\mathcal{F})$. Clearly, $\ell^p_m(\mathcal{F})$ is a linear space with respect to addition of sequences and scalar multiplication. We equip this space with the norm

$$\|(a_\nu)\|_{\ell^p_m(\mathcal{F})} := \|(\mathbf{a}_\nu)\|_{\ell^p(\mathcal{F})},$$

Now, if $(a_\nu)_{\nu \in \mathcal{F}} \in \ell^p_m(\mathcal{F})$, $0 < p < 1$, and Λ_k is any monotone realization of $\Lambda_k^*((\mathbf{a}_\nu)_{\nu \in \mathcal{F}})$, then the sets Λ_k are monotone and satisfy

$$\sum_{\nu \notin \Lambda_k} |a_\nu| \leq \sum_{\nu \notin \Lambda_k} \mathbf{a}_\nu \leq \|(a_\nu)\|_{\ell^p_m(\mathcal{F})} k^{-s}, \;\; s := \frac{1}{p} - 1. \tag{2.2}$$

Theorem 2.4 *Under the assumptions of Theorem 1.1, the sequence* $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ *belongs to* $\ell^p_m(\mathcal{F})$*.*

Proof: The result will follow from the estimates that underpin the proof of Theorem 1.1 given in [12]. A sequence $\rho = (\rho_j)_{j\geq 1}$ is said to be admissible of order δ if

$$\sum_{j\geq 1} \rho_j |\psi_j(x)| \leq \bar{a}(x) - \delta, \quad x \in D.$$

We denote by $\mathcal{A}_\delta$ the set of all δ-admissible sequences ρ for which $\rho_j \geq 1$, for all j. It was shown in [12] that for any $0 < \delta < r$

$$\|t_\nu\|_V \leq \frac{\|f\|_{V^*}}{\delta} \inf_{\rho\in\mathcal{A}_\delta} \rho^{-\nu}. \tag{2.3}$$

In particular, taking $\delta = \frac{r}{2}$,

$$\|t_\nu\|_V \leq b_\nu := 2C_0 \inf_{\rho\in\mathcal{A}_{\frac{r}{2}}} \rho^{-\nu}. \tag{2.4}$$

It was moreover shown that under the assumptions of Theorem 1.1 the sequence $(b_\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ which thus leads to the proof of Theorem 1.1. We now observe that the sequence b_ν is monotone because for any $\rho \in \mathcal{A}_{\frac{r}{2}}$

$$\mu \leq \nu \Rightarrow \rho^{-\nu} \leq \rho^{-\mu},$$

and thus

$$\mu \leq \nu \Rightarrow b_\nu \leq b_\mu.$$

Therefore, if $(\mathbf{a}_\nu)$ denotes the monotone majorant of the sequence $(\|t_\nu\|_V)$, we also find that

$$\mathbf{a}_\nu \leq b_\nu.$$

It follows that $\|(\|t_\nu\|_V)\|_{\ell^p_m(\mathcal{F})} \leq \|(b_\nu)\|_{\ell^p(\mathcal{F})} < \infty$. □

3 Recursive estimates

It was shown in [11] that the Taylor coefficients satisfy a recursion relation obtained by differentiating the variational formulation

$$\int_D a(x,y)\nabla u(x,y)\nabla v(x)dx = \int_D f(x)v(x)dx \quad v \in V,$$

at $y = 0$. Namely, we obtain by induction that $t_\nu \in V$ is the solution to the elliptic boundary value problem given in weak form by (see equations (4.6) and (4.10) of [11])

$$\int_D \bar{a}\nabla t_\nu \nabla v = -\sum_{j\,\text{s.t.}\,\nu_j\neq 0} \int_D \psi_j \nabla t_{\nu-e_j}\nabla v, \quad v \in V. \tag{3.1}$$

This recurrence allows one to compute all Taylor coefficients from the first coefficient $t_0 = u(0)$ corresponding to $\nu = 0$ which satisfies

$$\int_D \bar{a}\nabla t_0 \nabla v = \int_D fv, \quad v \in V. \tag{3.2}$$

In practice, these boundary value problems can only be solved approximately by space discretization, for example by the finite element method. We shall deal with this issue in §6 and assume for the moment that they can be solved exactly. For any monotone set of indices Λ, the recursion (3.1) determines the

Taylor coefficients $\{t_\nu \in V : \nu \in \Lambda\}$ uniquely; determining them requires the successive numerical solution of the "nominal" elliptic problems (3.2) with $\#(\Lambda)$ many right hand sides. In particular, then, for computing numerical approximations of the $(t_\nu)_{\nu\in\Lambda}$, *a discretized single, parameter-independent "nominal" elliptic problem* (3.2) *in the domain D must be solved* with $\#(\Lambda)$ many load cases. Since our adaptive algorithms will be based on the norms $\|t_\nu\|_{\bar{a}}$ of Taylor coefficients through the recursion (3.1), we introduce the abbreviated notation

$$\bar{t}_\nu := \|t_\nu\|_{\bar{a}}, \quad \nu \in \mathcal{F},$$

and the following quantities for bounding for the right hand side of (3.1)

$$d_{\mu,j} := \int_D |\psi_j||\nabla t_\mu|^2, \quad \mu \in \mathcal{F},\, j \geq 1.$$

We observe that $\mathbf{UEA}(r,R)$ implies for almost every $x \in D$

$$\sum_{j\geq 1} |\psi_j(x)| \leq \gamma \bar{a}(x) \tag{3.3}$$

with

$$\gamma = 1 - \frac{r}{R} < 1. \tag{3.4}$$

It follows that for any $\mu \in \mathcal{F}$, we have

$$\sum_{j\geq 1} d_{\mu,j} \leq \gamma \bar{t}_\mu^2. \tag{3.5}$$

Lemma 3.1 *Under assumption* $\mathbf{UEA}(r,R)$*, we have for any* $\nu \in \mathcal{F}$*,*

$$\bar{t}_\nu^2 \leq \alpha \sum_{j \text{ s.t. } \nu_j \neq 0} d_{\nu-e_j,j}, \tag{3.6}$$

with

$$\alpha := \frac{R}{R+r} < 1. \tag{3.7}$$

Proof: Taking $v = t_\nu$ in (3.1), we find that

$$\bar{t}_\nu^2 = -\sum_{j \text{ s.t. } \nu_j \neq 0} \int_D \psi_j \nabla t_{\nu-e_j} \nabla t_\nu, \tag{3.8}$$

and therefore

$$\bar{t}_\nu^2 \leq \frac{1}{2} \sum_{j \text{ s.t. } \nu_j \neq 0} \int_D |\psi_j|\, |\nabla t_{\nu-e_j}|^2 + \frac{1}{2} \sum_{j \text{ s.t. } \nu_j \neq 0} \int_D |\psi_j|\, |\nabla t_\nu|^2. \tag{3.9}$$

Using (3.3) in the second term of (3.9) gives

$$(1-\gamma/2)\bar{t}_\nu^2 \leq \frac{1}{2} \sum_{j \text{ s.t. } \nu_j \neq 0} \int_D |\psi_j|\, |\nabla t_{\nu-e_j}|^2,$$

from which we derive (3.6). □

For any set $\Lambda \subset \mathcal{F}$, we introduce

$$e(\Lambda) := \sum_{\nu\in\Lambda} \bar{t}_\nu^2, \qquad \sigma(\Lambda) := \sum_{\nu\in\mathcal{F}\setminus\Lambda} \bar{t}_\nu^2, \tag{3.10}$$

which is a measure of the energy of the Taylor coefficients on Λ and the energy on its complement respectively. Our next lemma shows that if Λ is a monotone set, the energy outside Λ is controlled by the energy on the margin $\mathcal{M} = \mathcal{M}(\Lambda)$.

Lemma 3.2 *Under assumption* (UEA)(r, R)*, we have for any monotone set* Λ *and its margin* $\mathcal{M}$*,*

$$\sigma(\Lambda) \leq \frac{1}{1-\delta} e(\mathcal{M}), \tag{3.11}$$

with

$$\delta = \frac{R-r}{R+r} < 1. \tag{3.12}$$

Proof: We first note that

$$\sigma(\Lambda) = e(\mathcal{M}) + \sigma(\tilde{\Lambda}), \tag{3.13}$$

where we have set $\tilde{\Lambda} := \Lambda \cup \mathcal{M}$. According to Lemma 3.1, we may write

$$\sigma(\tilde{\Lambda}) \leq \alpha \sum_{\nu \in \mathcal{F} \setminus \tilde{\Lambda}} \Big(\sum_{j \,\text{s.t.}\, \nu_j \neq 0} d_{\nu - e_j, j} \Big) \leq A + B, \tag{3.14}$$

where

$$A := \alpha \sum_{\nu \in \mathcal{F} \setminus \tilde{\Lambda}} \Big(\sum_{j \,\text{s.t.}\, \nu - e_j \in \mathcal{F} \setminus \tilde{\Lambda}} d_{\nu - e_j, j} \Big) = \alpha \sum_{\mu \in \mathcal{F} \setminus \tilde{\Lambda}} \Big(\sum_{j \,\text{s.t.}\, \mu + e_j \in \mathcal{F} \setminus \tilde{\Lambda}} d_{\mu, j} \Big),$$

and

$$B := \alpha \sum_{\nu \in \mathcal{F} \setminus \tilde{\Lambda}} \Big(\sum_{j \,\text{s.t.}\, \nu - e_j \in \tilde{\Lambda}} d_{\nu - e_j, j} \Big) = \alpha \sum_{\mu \in \mathcal{M}} \Big(\sum_{j \,\text{s.t.}\, \mu + e_j \in \mathcal{F} \setminus \tilde{\Lambda}} d_{\mu, j} \Big).$$

In this splitting, we have used the fact that if $\nu \in \mathcal{F} \setminus \tilde{\Lambda}$ and $\nu_j \neq 0$, we have either $\nu - e_j \in \mathcal{F} \setminus \tilde{\Lambda}$ or $\nu - e_j \in \mathcal{M}$. Using (3.5), we may control the first term A by

$$A \leq \alpha\gamma \sum_{\mu \in \mathcal{F} \setminus \tilde{\Lambda}} \bar{t}_\mu^2 = \alpha\gamma\sigma(\tilde{\Lambda}),$$

and by the same argument we obtain

$$B \leq \alpha\gamma e(\mathcal{M}).$$

Combining these estimates with (3.14), it follows that

$$(1 - \alpha\gamma)\sigma(\tilde{\Lambda}) \leq \alpha\gamma e(\mathcal{M}),$$

and thus by (3.13)

$$\sigma(\Lambda) \leq \Big(1 + \frac{\alpha\gamma}{1 - \alpha\gamma}\Big) e(\mathcal{M}),$$

which gives the final result. □

4 A bulk chasing algorithm

In this section, we introduce the notion of bulk chasing and show how this idea can be used to build an adaptive algorithm for generating a near optimal sequence of sets (Λ_n) in the sense of (1.7). We shall see that this algorithm is not numerically feasible but nevertheless will guide us in the construction of more practical algorithms in the sections that follow.

To begin the discussion, let us assume that we have a finite monotone set Λ with margin $\mathcal{M} = \mathcal{M}(\Lambda)$, for which we have already computed t_ν, $\nu \in \Lambda$. From this knowledge we can directly compute certain t_ν, $\nu \in \mathcal{M}$ from the recurrence (3.1). Namely, if $\mathcal{I}_1(\mathcal{M})$ is the set of $\nu \in \mathcal{M}$ such that each $\nu - e_j \in \Lambda$ whenever $\nu_j \geq 1$, then we can compute t_ν for all $\nu \in \mathcal{I}_1(\mathcal{M})$ since we already know each of the $t_{\nu - e_j}$ that occur in (3.1). We can then repeat this process and compute t_ν for any $\nu \in \mathcal{I}_2(\mathcal{M})$ where $\mathcal{I}_2(\mathcal{M})$ is the set of all $\nu \in \mathcal{M} \setminus \mathcal{I}_1(\mathcal{M})$ such that $\nu - e_j \in \Lambda \cup \mathcal{I}_1(\mathcal{M})$ whenever $\nu_j > 0$. Continuing in this way, we can compute all of the $t_\nu \in \mathcal{M}$. Notice that one only needs a finite number of the sets $\mathcal{I}_j(\mathcal{M})$ to exhaust $\mathcal{M}$ since Λ is finite.

For a fixed $0 < \theta < 1$, we consider the following (not yet practical) algorithm:

ALGORITHM 1

Define $\Lambda_0 := \{0\}$ and compute $t_0 := u(0)$ and $\bar{t}_0 := \|t_0\|_{\bar{a}}$. For $n = 0, 1, \cdots$ do the following:

- Given that Λ_n has been defined and $(t_\nu)_{\nu \in \Lambda_n}$ have been computed, we define $\mathcal{M}_n = \mathcal{M}(\Lambda_n)$ and compute t_ν, $\nu \in \mathcal{M}_n$, by using the recursion (3.1);
- Compute $\bar{t}_\nu$ for $\nu \in \mathcal{M}_n$;
- Find a smallest monotone set Λ_{n+1} such that $\Lambda_n \subset \Lambda_{n+1} \subset \Lambda_n \cup \mathcal{M}_n$ and $e(\Lambda_{n+1} \cap \mathcal{M}_n) \geq \theta e(\mathcal{M}_n)$;
- Go to step $n + 1$;

Note that requiring that Λ_{n+1} is monotone is equivalent to requiring that the update set $\Lambda_{n+1} \cap \mathcal{M}_n$ is monotone in $\mathcal{M}_n$.

This algorithm is not realistic for several reasons. First, we have already noticed that the margin $\mathcal{M}_n$ has infinite cardinality, and therefore there are infinitely many $\bar{t}_\nu$ to be computed which requires in principle solving infinitely many boundary value problems for the corresponding t_ν. We shall fix this problem in the next section §5. A second problem is that we can only solve the boundary value problems (3.1) approximately, for example using a finite element discretization. We analyze the additional error induced by this discretization in §6.

For the present, we remain with the above algorithm and prove its optimality. We first establish the following result on the decay of the energy of Taylor coefficients.

Theorem 4.1 *Under the assumptions of Theorem 1.1, we have $(\bar{t}_\nu)_{\nu \in \mathcal{F}} \in \ell^p_m(\mathcal{F})$ and the sets Λ_n satisfy*

$$\sigma(\Lambda_n) \leq C_1 \|(\bar{t}_\nu)\|^2_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n))^{-2t}, \quad t := \frac{1}{p} - \frac{1}{2}, \tag{4.1}$$

where C_1 only depends on (r, R, θ, t).

Proof: We have shown in Theorem 2.4 that $(\|t_\nu\|_V)_{\nu \in \mathcal{F}} \in \ell^p_m$. Since $\bar{t}_\nu \leq \sqrt{R}\|t_\nu\|_V$ for all $\nu \in \mathcal{F}$, it follows that $(\bar{t}_\nu)_{\nu \in \mathcal{F}} \in \ell^p_m(\mathcal{F})$ as stated.

To prove (4.1), we first observe that $\sigma(\Lambda_n)$ decreases geometrically. Indeed, we can write

$$\sigma(\Lambda_{n+1}) = \sigma(\Lambda_n) - e(\Lambda_{n+1} \cap \mathcal{M}_n) \leq \sigma(\Lambda_n) - \theta e(\mathcal{M}_n).$$

Since by (3.11), we have $e(\mathcal{M}_n) \geq (1 - \delta)\sigma(\Lambda_n)$, we obtain

$$\sigma(\Lambda_{n+1}) \leq \kappa \sigma(\Lambda_n), \tag{4.2}$$

with $\kappa := 1 - \theta(1 - \delta) = 1 - \frac{2\theta r}{R + r} < 1$.

We next control the cardinality of the updated set $\Lambda_{n+1} \cap \mathcal{M}_n$ by using the monotone majorants $\bar{\mathbf{t}}_\nu$ of the $\bar{t}_\nu$. Given any $\tau > 0$, we define k as the smallest integer such that

$$\|(\bar{t}_\nu)\|_{\ell^p_m(\mathcal{F})}(k+1)^{-t} \leq \tau \ .$$

Let $\mathcal{S}_k \subset \mathcal{M}_n$ be a monotone set in $\mathcal{M}_n$ corresponding to the k largest $\bar{\mathbf{t}}_\nu$ for $\nu \in \mathcal{M}_n$ (see Remark 2.3). From (1.5), we have

$$e(\mathcal{M}_n \setminus \mathcal{S}_k) = \sum_{\nu \in \mathcal{M}_n \setminus \mathcal{S}_k} \bar{t}_\nu^2 \leq \sum_{\nu \in \mathcal{M}_n \setminus \mathcal{S}_k} \bar{\mathbf{t}}_\nu^2 \leq \tau^2. \tag{4.3}$$

From the minimality of k, we also have

$$\#(\mathcal{S}_k) = k \leq \|(\bar{t}_\nu)\|^{1/t}_{\ell^p_m(\mathcal{F})} \tau^{-1/t} \ .$$

We now choose $\tau^2 := (1-\theta)e(\mathcal{M}_n)$ so that for this k we have from (4.3)

$$e(\mathcal{S}_k) \geq e(\mathcal{M}_n) - \tau^2 = \theta e(\mathcal{M}_n) \ .$$

Since $\mathcal{S}_k$ is a monotone set in $\mathcal{M}_n$, the set

$$\tilde{\Lambda}_{n+1} := \Lambda_n \cup \mathcal{S}_k, \tag{4.4}$$

is also monotone and satisfies the bulk property

$$e(\tilde{\Lambda}_{n+1} \cap \mathcal{M}_n) \geq \theta e(\mathcal{M}_n).$$

From the minimality of Λ_{n+1}, we thus have

$$\#(\Lambda_{n+1} \cap \mathcal{M}_n) \leq \#(\mathcal{S}_k) \leq \|(\bar{t}_\nu)\|^{1/t}_{\ell^p_m(\mathcal{F})} \tau^{-1/t} \ .$$

Using (3.11), it follows that

$$\#(\Lambda_{n+1} \cap \mathcal{M}_n) \leq [(1-\delta)(1-\theta)]^{-1/2t} \|(\bar{t}_\nu)\|^{1/t}_{\ell^p_m(\mathcal{F})} \sigma(\Lambda_n)^{-1/2t}. \tag{4.5}$$

Using the contraction property (4.2), we get $\sigma(\Lambda_k) \geq \kappa^{n-k}\sigma(\Lambda_n)$. We may now estimate the global cardinality of $\#(\Lambda_n)$ by

$$\begin{aligned} \#(\Lambda_n) \ &\leq \#(\Lambda_0) + \textstyle\sum_{k=0}^{n-1} \#(\Lambda_{k+1} \cap \mathcal{M}_k) \\ &\leq 1 + [(1-\delta)(1-\theta)]^{-1/2t} \|(\bar{t}_\nu)\|^{1/t}_{\ell^p_m(\mathcal{F})} \textstyle\sum_{k=0}^{n-1} \sigma(\Lambda_k)^{-1/2t} \\ &\leq 1 + [(1-\delta)(1-\theta)]^{-1/2t} \|(\bar{t}_\nu)\|^{1/t}_{\ell^p_m(\mathcal{F})} \sigma(\Lambda_n)^{-1/2t} \textstyle\sum_{k=0}^{n-1} \kappa^{(n-k)/2t} \\ &\leq 1 + C \|(\bar{t}_\nu)\|^{1/t}_{\ell^p_m(\mathcal{F})} \sigma(\Lambda_n)^{-1/2t} \end{aligned}$$

where $C := (\frac{\kappa}{(1-\delta)(1-\theta)}\Big)^{1/2t}(1-\kappa^{1/2t})^{-1}$. This last inequality can be rewritten as

$$\sigma(\Lambda_n) \leq C^{2t} \|(\bar{t}_\nu)\|^2_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n) - 1)^{-2t}.$$

If $\#(\Lambda_n) > 1$, we have established (4.1) with $C_1 := (2C)^{2t}$. If $\#(\Lambda_n) = 1$ then

$$\sigma(\Lambda_n) \leq \|(\bar{t}_\nu)\|^2_{\ell^2(\mathcal{F})} \leq \|(\bar{t}_\nu)\|^2_{\ell^p(\mathcal{F})} \leq \|(\bar{t}_\nu)\|^2_{\ell^p_m(\mathcal{F})},$$

and (4.1) also holds, with the constant C_1 only depending on r, R, θ and t. □

As a corollary, we also obtain the optimal decay of the ℓ^1 tail of the $\bar{t}_\nu$ and therefore of the error between u and its partial Taylor sum.

Corollary 4.2 *Under the assumptions of Theorem 1.1, we have $(\bar t_\nu)_{\nu\in\mathcal F}\in \ell^p_m(\mathcal F)$ and the sets Λ_n satisfy*

$$\sum_{\nu\in\mathcal F\setminus\Lambda_n}\bar t_\nu\le C_2\|(\bar t_\nu)\|_{\ell^p_m(\mathcal F)}(\#(\Lambda_n))^{-s},\ \ s:=\frac1p-1, \tag{4.6}$$

where $C_2:=1+\sqrt{C}_1$ with C_1 the constant in (4.1). Consequently we have

$$\sup_{y\in U}\|u(y)-\sum_{\nu\in\Lambda_n}t_\nu y^\nu\|_V\le\sum_{\nu\notin\Lambda_n}\|t_\nu\|_V\le\frac{1}{\sqrt r}\sum_{\nu\notin\Lambda_n}\bar t_\nu\le\frac{C_1}{\sqrt r}\|(\bar t_\nu)\|_{\ell^p_m(\mathcal F)}(\#(\Lambda_n))^{-s}. \tag{4.7}$$

Proof: Let $m:=\#(\Lambda_n)\ge 1$ and consider the set $\Lambda^*_m\subset\mathcal F$ corresponding to the m largest $\bar t_\nu$, we have

$$\begin{aligned}\textstyle\sum_{\nu\notin\Lambda_n}\bar t_\nu &= \textstyle\sum_{\nu\notin\Lambda^*_m}\bar t_\nu+\sum_{\nu\in\Lambda^*_m\setminus\Lambda_n}\bar t_\nu-\sum_{\nu\in\Lambda_n\setminus\Lambda^*_m}\bar t_\nu\\ &\le \textstyle\sum_{\nu\notin\Lambda^*_m}\bar t_\nu+\sum_{\nu\in\Lambda^*_m\setminus\Lambda_n}\bar t_\nu\\ &\le \|(\bar t_\nu)\|_{\ell^p(\mathcal F)}(m+1)^{-s}+m^{1/2}e(\Lambda^*_m\setminus\Lambda_n)^{1/2}\\ &\le (\|(\bar t_\nu)\|_{\ell^p(\mathcal F)}+\sqrt{C_1}\|(\bar t_\nu)\|_{\ell^p_m(\mathcal F)})(\#(\Lambda_n))^{-s}\\ &\le (1+\sqrt{C}_1)\|(\bar t_\nu)\|_{\ell^p_m(\mathcal F)}(\#(\Lambda_n))^{-s}\ ,\end{aligned}$$

where we have used both (1.5) and (4.1). This establishes (4.6), which implies (4.7) since $\|\cdot\|_V\le\frac{1}{\sqrt r}\|\cdot\|_{\bar a}$.□

5 A second algorithm

We now want to modify Algorithm 1 in order to restrict the computation of the t_ν to a finite subset of $\mathcal M_n$. In the modified algorithm, we set a target accuracy $\varepsilon>0$ and design the procedure in such a way that the algorithm terminates when $\sigma(\Lambda_n)\le\varepsilon$.

In order to restrict the margin $\mathcal M_n$ to a finite subset, we introduce a procedure SPARSE that has the following properties : if Λ is a finite monotone set and $\mathcal M$ is its infinite margin, and if $(\bar t_\nu)_{\nu\in\Lambda}$ are known, then for any $\eta>0$,

$$\mathcal N:=\mathrm{SPARSE}(\Lambda,(\bar t_\nu)_{\nu\in\Lambda},\eta),$$

is a finite subset of $\mathcal M$ which is monotone in $\mathcal M$ and such that $e(\mathcal M\setminus\mathcal N)\le\eta$.

There are several possible concrete realizations of this procedure. Here is a simple one. We define

$$\bar\psi_j:=\frac{\psi_j}{\bar a},$$

and choose $J>0$ large enough such that

$$\|\sum_{j>J}|\bar\psi_j|\|_{L^\infty(D)}\le\Big(\frac{\alpha e(\Lambda)}{1-\alpha\gamma}\Big)^{-1}\eta, \tag{5.1}$$

where α and γ are given by (3.4)and (3.7), and we define

$$\mathcal N:=\mathrm{SPARSE}(\Lambda,(\bar t_\nu)_{\nu\in\Lambda},\eta):=\{\nu\in\mathcal M\ ;\ \nu-e_j\in\Lambda\Rightarrow j\le J\}.$$

Clearly $\mathcal N$ is finite with $\#(\mathcal N)\le J\#(\Lambda)$.

Lemma 5.1 *With the above definition of $\mathcal N$, one has*

$$e(\mathcal M\setminus\mathcal N)=\sum_{\nu\in\mathcal M\setminus\mathcal N}\bar t_\nu^2\le\eta. \tag{5.2}$$

Proof: We proceed in a similar way to the proof of Lemma 3.2, by first writing

$$e(\mathcal{M} \setminus \mathcal{N}) \leq \alpha \sum_{\nu \in \mathcal{M} \setminus \mathcal{N}} \Big(\sum_{j \text{ s.t. } \nu_j \neq 0} d_{\nu - e_j, j} \Big) \leq A + B, \tag{5.3}$$

where now

$$A := \alpha \sum_{\nu \in \mathcal{M} \setminus \mathcal{N}} \Big(\sum_{j \text{ s.t. } \nu - e_j \in \mathcal{M} \setminus \mathcal{N}} d_{\nu - e_j, j} \Big) = \alpha \sum_{\mu \in \mathcal{M} \setminus \mathcal{N}} \Big(\sum_{j \text{ s.t. } \mu + e_j \in \mathcal{M} \setminus \mathcal{N}} d_{\mu, j} \Big),$$

and

$$B := \alpha \sum_{\nu \in \mathcal{M} \setminus \mathcal{N}} \Big(\sum_{j \text{ s.t. } \nu - e_j \notin \mathcal{M} \setminus \mathcal{N}} d_{\nu - e_j, j} \Big) = \alpha \sum_{\mu \in \Lambda \cup \mathcal{N}} \Big(\sum_{j \text{ s.t. } \mu + e_j \in \mathcal{M} \setminus \mathcal{N}} d_{\mu, j} \Big).$$

In this splitting, we have used the fact that if $\nu \in \mathcal{M} \setminus \mathcal{N}$ and $\nu_j \neq 0$, we have either $\nu - e_j \in \mathcal{M} \setminus \mathcal{N}$ or $\nu - e_j \in \Lambda \cup \mathcal{N}$. Using (3.5), we can bound A by

$$A \leq \alpha\gamma \sum_{\mu \in \mathcal{M} \setminus \mathcal{N}} \bar{t}_\mu^2 = \alpha\gamma e(\mathcal{M} \setminus \mathcal{N}).$$

To bound B, we first claim that for any $\mu \in \Lambda \cup \mathcal{N}$ such that $\mu + e_j \in \mathcal{M} \setminus \mathcal{N}$, we must have $\mu \in \Lambda$ and $j > J$. Indeed, since $\mu + e_j \in \mathcal{M} \setminus \mathcal{N}$, the definition of $\mathcal{N}$ guarantees that $\mu + e_j = \tilde{\nu} + e_k$ for some $\tilde{\nu} \in \Lambda$ and $k > J$. If $j = k$ we have our claim. If $j \neq k$ then necessarily $\tilde{\nu} - e_j \in \Lambda$ because of the monotonicity of Λ and therefore μ can be written as the sum of $\tilde{\nu} - e_j \in \Lambda$ and e_k, which means that μ is not in $\mathcal{N}$. Thus, we have verified our claim. From the claim, it follows that the only j's that may contribute in the summation inside B are such that $j > J$ and $\nu - e_j \in \Lambda$. Hence,

$$\begin{aligned} B &\leq \alpha \textstyle\sum_{\mu \in \Lambda} \sum_{j > J} d_{\mu, j} \\ &= \alpha \textstyle\sum_{\mu \in \Lambda} \int\limits_D (\sum_{j > J} |\psi_j|) |\nabla t_\nu|^2 \\ &= \alpha \textstyle\sum_{\mu \in \Lambda} \int\limits_D (\sum_{j > J} |\bar{\psi}_j|) \bar{a} |\nabla t_\nu|^2 \\ &\leq \alpha \| \textstyle\sum_{j > J} |\bar{\psi}_j| \|_{L^\infty} e(\Lambda) \leq (1 - \alpha\gamma)\eta. \end{aligned}$$

Combining the bounds for A and B with (5.3), we obtain

$$e(\mathcal{M} \setminus \mathcal{N}) \leq \frac{B}{1 - \alpha\gamma} \leq \eta,$$

as desired. □

For a fixed $0 < \theta < 1$ and target accuracy $\varepsilon > 0$, we now consider the following algorithm:

ALGORITHM 2

```
Define $\Lambda_0 := \{0\}$, compute $t_0 := u(0)$ and set $\eta_{0,0} := \bar{t}_0 = \|t_0\|_{\bar{a}}$;
For $n = 0, 1, \cdots$
  • Given $\Lambda_n$ and $(t_\nu)_{\nu \in \Lambda_n}$, define $\mathcal{M}_n := \mathcal{M}(\Lambda_n)$;
  • For $j = 0, 1, \cdots$
      – Define $\eta_{n,j} := 2^{-j}\eta_n$ and $\mathcal{M}_{n,j} := \mathrm{SPARSE}(\Lambda_n, (\bar{t}_\nu)_{\nu \in \Lambda_n}, \eta_{n,j})$;
      – Compute $\bar{t}_\nu$ for $\nu \in \mathcal{M}_{n,j}$ and compute $e(\mathcal{M}_{n,j})$;
```

- If $e(\mathcal{M}_{n,j}) + \eta_{n,j} \le (1-\delta)\varepsilon$, with δ as in (3.12), then terminate the Algorithm and output the set $\Lambda(\varepsilon) := \Lambda_n$;
- Else if $e(\mathcal{M}_{n,j}) < \frac{4-2\theta}{1-\theta}\eta_{n,j}$, then go directly to step $j+1$;
- Else if $e(\mathcal{M}_{n,j}) \ge \frac{4-2\theta}{1-\theta}\eta_{n,j}$, then terminate the inner loop in j, and define $\eta_{n+1} := \eta_{n,j}$ and $\Lambda_{n+1} := \mathcal{S}_{n,j} \cup \Lambda_n$, where $\mathcal{S}_{n,j} \subset \mathcal{M}_{n,j}$ is the smallest monotone set in $\mathcal{M}_{n,j}$ such that
$$\sum_{\nu\in\mathcal{M}_{n,j}\setminus\mathcal{S}_{n,j}} \bar{t}_\nu^2 \le (1-\theta)e(\mathcal{M}_{n,j}) - (2-\theta)\eta_{n,j}.$$

- Compute t_ν for $\nu \in \Lambda_{n+1}$ using (3.1);
- Go to step $n+1$;

Theorem 5.2 *Under the assumptions of Theorem 1.1 with $p \le 1$, Algorithm 2 terminates for a finite value n^* and outputs a set $\Lambda := \Lambda(\varepsilon) := \Lambda_{n^*}$ which satisfies*
$$\sigma(\Lambda) \le \varepsilon. \tag{5.4}$$
Moreover, one has
$$\sigma(\Lambda) \le \bar{C}_1 \|(\bar{t}_\nu)\|^2_{\ell^p_m(\mathcal{F})}(\#(\Lambda))^{-2t}, \quad t := \frac{1}{p} - \frac{1}{2}, \tag{5.5}$$
and
$$\sup_{y\in U} \|u(y) - \sum_{\nu\in\Lambda} t_\nu y^\nu\|_V \le \bar{C}_2 \|(\bar{t}_\nu)\|_{\ell^p_m(\mathcal{F})}(\#(\Lambda))^{-s}, \quad s := \frac{1}{p} - 1, \tag{5.6}$$
where the constants $\bar{C}_1$ and $\bar{C}_2 := \frac{1+\sqrt{\bar{C}_1}}{\sqrt{r}}$ only depend on (r, R, θ, t).

Proof: We first claim that for each n the inner j loop terminates for some j. To see this, we note that this loop advances j only when $e(\mathcal{M}_{n,j}) < \frac{4-2\theta}{1-\theta}\eta_{n,j}$ and $e(\mathcal{M}_{n,j}) > (1-\delta)\varepsilon - \eta_{n,j}$. This cannot happen when j is large because the $\eta_{n,j}$ tend to zero with increasing j.

Next, we need to check that for the output j of the inner loop, we are able to determine the index set $\mathcal{S}_{n,j}$. To see this, we note that for the output j of the inner loop, we must have $e(\mathcal{M}_{n,j}) > \frac{4-2\theta}{1-\theta}\eta_{n,j}$ and therefore
$$(1-\theta)e(\mathcal{M}_{n,j}) - (2-\theta)\eta_{n,j} > (2-\theta)\eta_{n,j} > 0.$$
So, we are indeed able to find the smallest monotone $\mathcal{S}_{n,j} \subset \mathcal{M}_{n,j}$ such that
$$\tilde{e}(\mathcal{M}_n \setminus \mathcal{S}_{n,j}) \le (1-\theta)e(\mathcal{M}_{n,j}) - (2-\theta)\eta_{n,j}.$$
Since $\mathcal{M}_{n,j} \subset \mathcal{M}_n$, we have
$$e(\mathcal{M}_n \setminus \mathcal{S}_{n,j}) \le (1-\theta)e(\mathcal{M}_{n,j}) \le (1-\theta)e(\mathcal{M}_n). \tag{5.7}$$
Next we have to check that the outer n loop terminates. To see this we note that from (5.7) and the same reasoning as in the proof of Theorem 4.1, the contraction property
$$\sigma(\Lambda_{n+1}) \le \kappa\sigma(\Lambda_n), \tag{5.8}$$
holds whenever the set Λ_{n+1} is created by the algorithm and $0 < \kappa < 1$ is the constant of (4.2). Obviously, this contraction property shows that for any given $c > 0$, if the algorithm did not terminate for n large

enough we must have $e(\mathcal{M}_n) \le c$ because $e(\mathcal{M}_n) \le \sigma(\Lambda_n)$. Here we take $c = \frac{(4-2\theta)(1-\delta)\varepsilon}{5-3\theta}$. If such a value of n has been reached by the algorithm and if Λ_n is not selected during the inner loop, this means that for the last value of j in this inner loop, we have

$$\eta_{n,j} \le \frac{1-\theta}{4-2\theta} e(\mathcal{M}_{n,j}),$$

and thus

$$e(\mathcal{M}_{n,j}) + \eta_{n,j} \le \frac{5-3\theta}{4-2\theta} e(\mathcal{M}_{n,j}) \le \frac{5-3\theta}{4-2\theta} e(\mathcal{M}_n) \le (1-\delta)\varepsilon.$$

Therefore, the algorithm terminates.

Let n^* be the terminal value of n and let j^* be the terminal value of j for the inner j loop applied to this value $n = n^*$. We have $e(\mathcal{M}_{n^*,j^*}) + \eta_{n^*,j^*} \le (1-\delta)\varepsilon$ and thus

$$\sigma(\Lambda_{n^*}) \le (1-\delta)^{-1} e(\mathcal{M}_{n^*}) \le (1-\delta)^{-1}(e(\mathcal{M}_{n^*,j^*}) + \eta_{n^*,j^*}) \le \varepsilon.$$

Therefore, (5.4) holds for the generated set $\Lambda = \Lambda_{n^*}$.

We shall next prove (5.5) by using an argument similar to that used in the proof of Theorem 4.1. For any $n < n^*$, we will bound the cardinality of the update set $\mathcal{S}_{n,j}$ where $j = j(n)$ is the terminal value of j for this n. Fix such a pair n, j and define $\tau^2 = \tau_n^2 := (1-\theta)e(\mathcal{M}_{n,j}) - (2-\theta)\eta_{n,j}$. We define k as the smallest integer such that

$$\|(\bar{\mathbf{t}}_\nu)\|_{\ell^p(\mathcal{F})}(k+1)^{-t} = \|(\bar{t}_\nu)\|_{\ell^p_m(\mathcal{F})}(k+1)^{-t} \le \tau.$$

From (1.5), we see that the set $\Gamma_k \subset \mathcal{M}_n$ corresponding to the k largest $\bar{\mathbf{t}}_\nu$ for $\nu \in \mathcal{M}_{n,j}$, satisfies

$$\sum_{\nu \in \mathcal{M}_n \setminus \Gamma_k} \bar{t}_\nu^2 \le \sum_{\nu \in \mathcal{M}_n \setminus \Gamma_k} \bar{\mathbf{t}}_\nu^2 \le \tau^2 \ .$$

From the minimality of k, we also have

$$\#(\Gamma_k) = k \le \|(\bar{t}_\nu)\|_{\ell^p_m(\mathcal{F})}^{1/t} \tau^{-1/t}.$$

The set Γ_k can be taken monotone in $\mathcal{M}_{n,j}$ (see Remark 2.3), and therefore from the minimality of $\mathcal{S}_{n,j}$, we have

$$\#(\mathcal{S}_{n,j}) \le \#(\Gamma_k) \le \|(\bar{t}_\nu)\|_{\ell^p_m(\mathcal{F})}^{1/t} \tau^{-1/t}.$$

We can now estimate the global cardinality of $\#(\Lambda) = \#(\Lambda_{n^*})$ by

$$\#(\Lambda) \le \#(\Lambda_0) + \sum_{n=0}^{n^*-1} \#(\mathcal{S}_{n,j(n)}) \le 1 + \|(\bar{t}_\nu)\|_{\ell^p_m(\mathcal{F})}^{1/t} \sum_{k=0}^{n^*-1} \tau_n^{-1/t}. \tag{5.9}$$

We need a lower bound for the τ_n which is given by

$$\begin{aligned} \tau_n^2 = (1-\theta)e(\mathcal{M}_{n,j}) - (2-\theta)\eta_{n,j} \quad &= \tfrac{1}{3}(1-\theta)e(\mathcal{M}_{n,j}) + \tfrac{2}{3}(1-\theta)e(\mathcal{M}_{n,j}) - (2-\theta)\eta_{n,j} \\ &\ge \tfrac{1}{3}(1-\theta)e(\mathcal{M}_{n,j}) + \tfrac{2}{3}(4-2\theta)\eta_{n,j} - (2-\theta)\eta_{n,j} \\ &= \tfrac{1}{3}(1-\theta)e(\mathcal{M}_{n,j}) + \tfrac{1}{3}(2-\theta)\eta_{n,j} \\ &\ge \tfrac{1-\theta}{3}(e(\mathcal{M}_{n,j}) + \eta_{n,j}) \\ &\ge \tfrac{1-\theta}{3} e(\mathcal{M}_n) \ge (1-\delta)\Big(\tfrac{1-\theta}{3}\Big)\tilde{\sigma}(\Lambda_n) =: \bar{C}_2^{-1}\sigma(\Lambda_n). \end{aligned}$$

Here the last inequality follows from Lemma 3.2 and the next to last inequality follows from the fact that $e(\mathcal{M}_n \setminus \mathcal{M}_{n,j}) \leq \eta_{n,j}$ by virtue of (5.2). If we place this lower bound for τ_n into (5.9), we obtain

$$\begin{aligned}\#(\Lambda) &\leq 1 + \bar{C}_2^{1/t} \|(\bar{t}_\nu)\|_{\ell_m^p(\mathcal{F})}^{1/t} \textstyle\sum_{n=0}^{n^*-1} \sigma(\Lambda_n)^{-1/2t} \\ &\leq 1 + \bar{C}_2^{1/t} \|(\bar{t}_\nu)\|_{\ell_m^p(\mathcal{F})}^{1/t} \sigma(\Lambda)^{-1/2t} \textstyle\sum_{k=0}^{n^*-1} \kappa^{(n^*-k)/2t} \\ &\leq 1 + C \|(\bar{t}_\nu)\|_{\ell_m^p}^{1/t} \sigma(\Lambda)^{-1/2t},\end{aligned}$$

where $C := \bar{C}_2^t (1-\kappa^{1/2t})^{-1}$. This last inequality can be rewritten as

$$\sigma(\Lambda) \leq \bar{C}_1 \|(\bar{t}_\nu)\|_{\ell_m^p(\mathcal{F})}^2 (\#(\Lambda))^{-2t},$$

in the same way we argued to complete the proof of Theorem 4.1. Hence, we have proved (5.5). Finally, (5.6) follows from (5.5) in the same way we have derived Corollary 4.2 from Theorem 4.1. □

Although Algorithm 2 algorithm meets the benchmark of the optimal rate (1.7) under the minimal assumptions of Theorem 1.1, a closer inspection shows that it is not completely optimal from a computational point of view. Indeed, consider the number $B = B(\varepsilon) = B_{n^*}$ of boundary value problems which have actually been solved in order to compute the functions t_ν for ν in the final set $\Lambda = \Lambda(\varepsilon) = \Lambda_{n^*}$. Ideally, we would hope that this number is not much larger than the cardinality of Λ, so that we may actually retrieve the convergence estimates (5.5) and (5.6) in terms of B instead of $\#(\Lambda)$.

However, the number B involves the size of the restricted margin which is produced by the procedure SPARSE, and which might in principle be substantially larger than the set that is finally selected by the bulk search. Retrieving the same convergence rate in terms of B would actually require that when the accuracy η prescribed in SPARSE is of the same order as the current accuracy $e(\Lambda)$, then the cardinality of the produced set $\mathcal{N}$ should be bounded by the optimal rate

$$\#(\mathcal{N}) \leq C \|(\bar{t}_\nu)\|_{\ell_m^p(\mathcal{F})}^{1/t} \eta^{-1/2t}. \tag{5.10}$$

A brief inspection seems to indicate that only a lower rate is achieved by our SPARSE procedure: on the one hand we know that

$$\#(\mathcal{N}) \leq J\#(\Lambda),$$

and that the set Λ has its cardinality optimally controlled by $\eta^{-1/2t}$, and on the other hand the number J that ensures (5.1) is of the order $\eta^{-1/s}$ where $s = \frac{1}{p} - 1 = t - \frac{1}{2}$. Therefore $\eta^{-1/2t}$ in (5.10) is a-priori replaced by the non-optimal rate $\eta^{-1/(2t^2-t)}$

In order to remedy this defect, one would need to design more elaborate realizations of SPARSE in order to obtain a set $\mathcal{N}$ of smaller, hopefully optimal, cardinality. One option that could lead to such a SPARSE procedure would be to make use of the available *a-priori bounds* on the $\|t_\nu\|_V$ such as such as (2.3) and (2.4) in order to control the energy outside of the set $\mathcal{N}$. We do not embark in this direction here. Another option for lowering the CPU cost, which appears to work quite well in practice yet without a complete theoretical justification, will be proposed in §7.

6 Space discretization

The boundary value problems that recursively give the coefficients e_ν cannot be solved exactly. Instead, we would use a Galerkin method in a finite dimensional space $V_h \subset V$, typically a finite element space although this is not crucial in the present analysis which would also apply to spectral or wavelet discretization. We shall show in this section that it is possible to choose the *same* space V_h to approximate *all* e_ν and still retain the performance of Algorithm 1 and Algorithm 2.

For the purpose of simplicity, we consider here the situation where the same spatial discretization is used for all ν. Yet, the analysis in §8 of [11] reveals that substantial computational gain may be expected if the spatial discretization is allowed to vary with ν (typically, coarser discretizations should be used for the computation of smaller Taylor coefficients). The possibility of adaptively choosing the approximation space parameter h depending on ν should also be explored but requires a more involved analysis. A future objective is therefore to design a solution algorithm that adaptively monitors the spatial resolution as new coefficients are being computed.

We define the Finite Element approximation $u_h(y) \in V_h$ as the solution to

$$\int_D a(x,y)\nabla u_h(y)\nabla v_h = \int_D f v_h \qquad \forall v_h \in V_h \ . \tag{6.1}$$

By assumption $\mathbf{UEA}(r,R)$, for any closed subspace $V_h \subset V$ the Finite Element approximation is uniquely defined and the analysis in [11] and [12] and all results of the present paper also apply to the discretized problem.

In particular, for every $h > 0$, the Finite Element approximation $u_h(y) \in V_h$ can be represented as a convergent Taylor expansion about $y = 0$ i.e., $u_h(y) = \sum_{\nu\in\mathcal{F}} t_{\nu,h} y^\nu$, where

$$t_{\nu,h} := \frac{1}{\nu!}\partial^\nu u_h(0) \in V_h \ .$$

Moreover, the $\|t_{\nu,h}\|_V$ can be estimated by the same bound $b_\nu := 2C_0 \inf_{\rho\in\mathcal{A}_r} \rho^{-\nu}$ as the $\|t_\nu\|_V$, leading to a result similar to Theorem 2.4.

Theorem 6.1 *Under the assumptions of Theorem 1.1, the sequence $(\|t_{\nu,h}\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$. Moreover $\|(\|t_{\nu,h}\|_V)\|_{\ell^p(\mathcal{F})}$ is bounded independent of h.*

The coefficients $t_{\nu,h}$ can be computed recursively by solving linear systems corresponding to the space-discretized boundary value problems

$$\int_D \bar{a}\nabla t_{\nu,h}\nabla v_h = -\sum_{j\,\text{s.t.}\,\nu_j\neq 0}\int_D \psi_j \nabla t_{\nu-e_j,h}\nabla v_h \quad \forall v_h \in V_h. \tag{6.2}$$

For the approximate Taylor coefficients, we introduce once more their energies as $\bar{t}_{\nu,h} := \|t_{\nu,h}\|_{\bar{a}}$. We may define $e_h(\Lambda)$ and $\sigma_h(\Lambda)$ and apply Algorithms 1 or 2 in a similar way, by simply replacing t_ν and $\bar{t}_\nu$ by $t_{\nu,h}$ and $\bar{t}_{\nu,h}$. For these algorithms, we obtain convergence results by the exact same approach as without space discretization.

Theorem 6.2 *Under the assumptions of Theorem 1.1 with $p < 1$, the application of each of the Algorithms 1 or 2 in the space discretized setting yields a sequence of sets (Λ_n) such that*

$$\sigma_h(\Lambda_n) \le C_1 \|(\bar{t}_{\nu,h})\|_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n))^{-2t}, \ \ t := \frac{1}{p} - \frac{1}{2}, \tag{6.3}$$

and

$$\sum_{\nu\notin\Lambda_n} e_{\nu,h} \le C_2 \|(\bar{t}_{\nu,h})\|_{\ell^p_m} (\#(\Lambda_n))^{-s}, \ \ s := \frac{1}{p} - 1, \tag{6.4}$$

where C_1 and C_2 are as in the continuous setting (depending on r, R, θ and on t, but being independent of h).

Consequently we have

$$\sup_{y\in U}\|u_h(y)-\sum_{\nu\in\Lambda_n}t_{\nu,h}y^\nu\|_V\leq\sum_{\nu\notin\Lambda_n}\|t_{\nu,h}\|_V\leq\frac{C_1}{\sqrt{r}}\|(\bar{t}_{\nu,h})\|_{\ell^p_m}(\#(\Lambda_n))^{-s}. \tag{6.5}$$

We finally quantify the space discretization error. The well-known theory of finite elements tells us that the rate of convergence of $\|u(y)-u_h(y)\|_V$ in terms of the decay of h is controlled by the smoothness of u in the scale of the H^s Sobolev space and the order of the finite element spaces V_h which are employed. For example, when using Lagrange finite elements of order k, we have

$$\sup_{y\in U}\|u(y)-u_h(h)\|_V\leq C_3h^r\sup_{y\in U}\|u(y)\|_{H^{1+r}(D)}, \tag{6.6}$$

for all $r\leq k$. This leads to the following result.

Corollary 6.3 *Under the assumptions of Theorem 1.1 and assuming that* $\sup_{y\in U}|u(y)|_{H^{1+r}}<\infty$ *and that we use Lagrange finite elements of order* $k\geq r$*, then applying Algorithms 1 or 2 in the space discretized setting, we obtain*

$$\sup_{y\in U}\|u(y)-\sum_{\nu\in\Lambda_n}t_{\nu,h}y^\nu\|_V\leq C_3h^r\sup_{y\in U}\|u(y)\|_{H^{1+r}(D)}+\frac{C_1}{\sqrt{r}}\|(\bar{t}_{\nu,h})\|_{\ell^p_m(\mathcal{F})}(\#(\Lambda_n))^{-s}. \tag{6.7}$$

The largest value of r for which $\sup_{y\in U}|u(y)|_{H^{1+r}}<\infty$ is determined by

- the smoothness of the right hand side f,
- the smoothness of the diffusion coefficient a,
- the smoothness of the boundary of D.

As an example, for $f\in L^2(D)$ and coefficients $a(x,y)$ which satisfy

$$\sup_{y\in U}\|a(\cdot,y)\|_{W^{1,\infty}(D)}<\infty, \tag{6.8}$$

we note that (1.1) implies that the solution $u(y)$ satisfies the Poisson equation

$$-\Delta u(y)=\frac{1}{a}[f-\nabla a\cdot\nabla u(y)] \quad \text{in} \quad D\ , \qquad u(y)|_{\partial D}=0\ . \tag{6.9}$$

Therefore, for every $y\in U$ the solution $u(y)$ belongs to the space

$$W=\left\{v\in V:\Delta v\in L^2(D)\right\}\ .$$

If the domain D is convex, it is well known $W=H^2(D)\cap H^1_0(D)$. For more general Lipschitz domains, it is also known that $W=H^{1+r}(D)\cap H^1_0(D)$ for some $\frac{1}{2}\leq r\leq 1$. We refer to [18] for a general treatment of elliptic problems on non-smooth domains.

In the numerical experiment that follow, we actually deal with coefficients $a(x,y)$ which are piecewise constant on a partition of $D=[0,1]^2$ into fixed sub-squares independent of y. Such coefficients obviously do not satisfy (6.8), however regularity results are also known in this setting and give that the solution $u(y)$ belong to $H^{1+r}(D)\cap H^1_0(D)$ for some $0<r\leq\frac{1}{2}$ that depends on the maximal contrast R/r, see for example [3].

7 Numerical Experiments

In this section, we study the numerical performance of the algorithm proposed in this paper. In this algorithm, the choice of Λ_n is made adaptively and it is based on a bulk search procedure. In particular, we want to compare this choice of Λ_n, with non-adaptive choices. We also study some variants using other adaptive strategies.

The algorithms that we analyzed in the previous sections were formulated regardless of wether the dimension of y is finite or infinite. In the present numerical test, we use a parameter vector $y = (y_j)_{j=1,\cdots,d}$, of dimension 64, i.e. $y \in [-1,1]^{64}$. More precisely, we consider the following numerical test on the unit square $D := [0,1] \times [0,1]$:

$$-\text{div}(a\nabla u) = f \text{ in } D, \ \ u = 0 \text{ on } \partial D,$$

where for illustration purposes we take $f(x_1, x_2) := x_1 x_2$. We partition D into 64 squares D_j of equal shape and consider a diffusion coefficient that is piecewise constant on each subdomain:

$$a(x,y) = \bar{a} + \sum_{j=1}^{64} y_j \psi_j, \ \text{ where } \ \bar{a} = 1 \ \text{ and } \ \psi_j = \alpha_j \chi_{D_j}. \tag{7.1}$$

Since in this case the ψ_j have disjoint supports, the uniform ellipticity assumption simply means that the weights $\alpha_j = \|\psi_j\|_{L^\infty(D)}$ are all strictly less than 1. To study the consistency of the numerical results with our theory, we also need that the sequence α_j has some decay, since in the case of an infinite sequence we require that $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$ is summable. In our numerical test we take in (7.1)

$$\alpha_j = \frac{0.9}{j^3}. \tag{7.2}$$

The uniform ellipticity assumption $\mathbf{UEA}(r,R)$ therefore holds with $r = 0.1$.

As mentioned above, we use one fixed finite element space for the spatial discretization of all active Taylor coefficients. Therefore, for the different strategies of building the coefficients sets Λ_n, we actually study the decay of Taylor expansion error *for the finite element solution*

$$\sup_{y\in U} \|u_h(y) - \sum_{\nu\in\Lambda_n} t_{\nu,h} y^\nu\|_V, \tag{7.3}$$

as $\#(\Lambda_n)$ grows, bearing in mind that the finite element discretization induces an additional source of error $\sup_{y\in U} \|u(y) - u_h(y)\|_V$ which can be bounded according to (6.6).

For the generation of the sets Λ_n of "active" Taylor coefficients, we compare three non-adaptive strategies that are based on a-priori choices of the sets Λ_n, and three adaptive strategies (in particular Algorithm 1) that exploit the results of earlier computations. For the sake of notational simplicity, we describe these algorithms without the additional finite element discretization setting, therefore using the notation t_ν instead of $t_{\nu,h}$. Their adaptation in the finite element setting is of course straightforward.

Non-adaptive strategies:

- **Algorithm QN:** For $n \geq 0$, we take $\Lambda_n = \{\nu \in \mathcal{F}, \text{ s.t } \max(\nu_j) \leq n\}$. Therefore $\text{Span}\{y \mapsto y^\nu \ ; \ \nu \in \Lambda_n\}$ is the space $\mathbb{Q}_n$ of polynomials of degree at most n in each variable. The dimension of this space $\#\Lambda_n = (n+1)^d$ grows exponentially with the dimension d of y, reflecting the curse of dimensionality.
- **Algorithm PN:** For $n \geq 0$, we take $\Lambda_n = \{\nu \in \mathcal{F}, \text{ s.t } |\nu| \leq n\}$. Therefore $\text{Span}\{y \mapsto y^\nu \ ; \ \nu \in \Lambda_n\}$ is the space $\mathbb{P}_n$ of polynomials of total degree at most n. The dimension of this space $\#\Lambda_n = \binom{n+d}{n}$, although smaller than that of $\mathbb{Q}_n$ by an order $d!$ still grows exponentially with d.

- **Algorithm LE (Largest Estimates):** as explained in the proof of Theorem 2.4, there are available estimates such as (2.3) and (2.4) for $\|t_\nu\|_V$. It is therefore natural to choose for Λ_n the set of indices corresponding to the n largest of these estimates, for example the n largest

  $$e_\nu := \frac{\|f\|_{V^*}}{\delta} \inf_{\rho \in \mathcal{A}_\delta} \rho^{-\nu},$$

  for some given $0 < \delta < r$. As already explained, such sets are monotone by construction. In practice, it is not always simple to compute the exact value of the infimum in the above definition of e_ν. However, this problem has a simple solution in the case where the ψ_j's have disjoint supports since all ρ_j can be optimized separately. For our problem this easily leads to the solution

  $$\rho_j^* = \frac{1-\delta}{\alpha_j},$$

  and therefore

  $$e_\nu := \frac{\|f\|_{V^*}}{\delta} \prod_{j=1}^d \Big(\frac{\alpha_j}{1-\delta}\Big)^{\nu_j}.$$

  The set Λ_n may also be viewed as the set of those ν such that e_ν exceeds a certain threshold $t = t(n) > 0$ that decreases with n, and are therefore of the form

  $$\Lambda_n := \Big\{\nu \ ; \ \sum_{j=1}^d a_j \nu_j \leq \Theta(n)\Big\} \ \text{ with } \ a_j := -\log\Big(\frac{\alpha_j}{1-\delta}\Big) \ \text{ and } \ \Theta(n) = -\log\Big(\frac{t(n)\delta}{\|f\|_{V^*}}\Big).$$

  Note that if all α_j - and therefore a_j - were equal, then this would give the same a-priori choice as the previously described Algorithm PN. In our case, the a_j decrease with j, resulting in some anisotropy in the sets Λ_n: higher polynomial degrees are expected for small values of j which represent the most "active" variables. A similar choice of polynomial space was studied in [25] for collocation methods.

  In our numerical tests, we have used the value $\delta = \frac{r}{2} = 0.05$. One could use an even sharper a-priori estimate on the $\|t_\nu\|_V$ by taking the infimum of e_ν also over all $\delta \in]0, r[$. This leads to a similar estimate but now with δ depending on ν according to $\delta = \min(r, \frac{1}{1+|\nu|})$.

Adaptive strategies:

- **Algorithm BS (Bulk Search):** This is simply Algorithm 1 based on the bulk search procedure as proposed in §4, and applied under the given finite element discretization. Since we work in finite dimension, it is in theory possible to apply this algorithm without the need to restrict the margin as it is done in Algorithm 2. In our numerical tests, we have built the new set Λ_{n+1} by calculating the monotone majorant of the sequence $\bar{t}_\nu := \|t_\nu\|_{\bar{a}}$ for $\nu \in \mathcal{M}_n$ (extended by 0 outside of $\mathcal{M}_n$), and then by adding to Λ_n the smallest set $\mathcal{S}_k$ corresponding to the k largest $\bar{\mathbf{t}}_\nu$ for which $e(\mathcal{S}_k) \geq \theta e(\mathcal{M}_n)$. The new set Λ_{n+1} is monotone by construction. However, it is not exactly the smallest monotone set such that $e(\mathcal{M}_n \cap \Lambda_{n+1}) \geq \theta e(\mathcal{M}_n)$. The construction of this optimal monotone set by a fast algorithm is still an open problem to us. In our numerical test, we have used the value $\theta = 0.2$ for the bulk parameter (we observed that the error curves are almost identical when θ ranges in $[0.05, 0.95]$).

- **Algorithm LN (Largest Neighbor):** Although Algorithm 1 may be performed in the finite dimensional context, the size of the current margin $\mathcal{M}_n$ relative to the size of the current set Λ_n becomes a source of computational slow-down as d grows. An alternate strategy is to only consider the *reduced margin*

$$\mathcal{I}_1(\Lambda_n) := \{\nu \notin \Lambda_n \ \ ; \ \ \nu_j \neq 0 \Rightarrow \nu - e_j \in \Lambda_n\},$$

which are the indices in $\mathcal{M}_n$ for which the Taylor coefficients t_ν can directly be computed from those indexed by Λ_n. We then define

$$\Lambda_{n+1} = \Lambda_n \cup \{\nu^*\},$$

where

$$\nu^* := \text{Argmax}_{\nu \in \mathcal{I}_1(\Lambda_n)} \bar{t}_\nu.$$

The intuition for considering such a strategy is that if the sequence $(\bar{t}_\nu)_{\nu \in \mathcal{F}}$ were monotone, then this would select the $\bar{t}_\nu$ in decreasing order. The potential pay-off is that the reduced margin is much smaller than $\mathcal{M}_n$ (in particular, it is easy to check that at most d boundary value problems need to be solved at each iteration). As we shall see, this strategy gives excellent results, although we have no proof similar to Algorithm 1 that it performs optimally in the sense of convergence rates.

- **Algorithm LNE (Largest Neighbor Estimate):** In order to save further computational cost, we can use majorants of $\bar{t}_\nu$ in order to decide on the new set Λ_{n+1}. From (3.7), one straightforward upper estimate for $\bar{t}_\nu$ is
$$\bar{t}_\nu \leq N_\nu := \Big(\alpha \sum_{j \text{ s.t. } \nu_j \neq 0} \|\bar{\psi}_j\|_{L^\infty(D)} \bar{t}^2_{\nu - e_j}\Big)^{\frac{1}{2}} \tag{7.4}$$
One can then construct the new set Λ_{n+1} as in the previous Algorithm LN, by using N_ν instead of $\bar{t}_\nu$. The saving comes from the fact that computing N_ν is much cheaper than computing t_ν.

We have compared the various strategies using 4 choices of finite element spaces based on uniform triangulations of D obtained by splitting each element of a square mesh into two triangles: (i) 8 × 8 squares and $\mathbb{P}_1$ finite elements ($\dim(V_h) = 49$), (ii) 16 × 16 squares and $\mathbb{P}_1$ finite elements ($\dim(V_h) = 225$) (iii) 16 × 16 squares and $\mathbb{P}_2$ finite elements ($\dim(V_h) = 961$), (iv) 32 × 32 squares and $\mathbb{P}_1$ finite elements ($\dim(V_h) = 961$). We display on Figure 7.1 the error curves for the six strategies described above for the generation of the sets Λ_n. These error curves represent the supremum error (7.3) (estimated by taking the supremum over a random choice of 100 values of y) as a function of $\#(\Lambda_n)$). Note that for certain strategies, such as PN, QN and BS, the number $\#(\Lambda_n)$ does not grow by 1 at each iteration and therefore only takes a few integer values. In such cases, we obtain all intermediate values for the error curves by filling the intermediates indices in $\Lambda_{n+1} \setminus \Lambda_n$ by lexicographic order.

We also indicate for each choice of finite element space an estimate of the FE error $\sup_{y \in U} \|u(y) - u_h(y)\|_V$. This estimate is done by replacing $u(y)$ by a finite element solution on a very fine mesh obtained from 256×256 squares and taking the supremum over the same random choice of 100 values of y.

We record three major observations about the error curves.

- First, not much difference in the error curves is observed as we modify the spatial discretization, once it is finer than 8 × 8. In fact, a closer inspection also show that the sets Λ_n selected by the adaptive algorithms change very little as we modify the spatial discretization. This suggest that the same sets and error curves would be obtained if there there were no spatial discretization at all, i.e. if we were computing the t_ν by exactly solving the boundary value problems (3.1). In particular, the portion of the error curves which is below the value of the finite element error is still relevant to us, since this portion does not seem to change as this error is diminished.

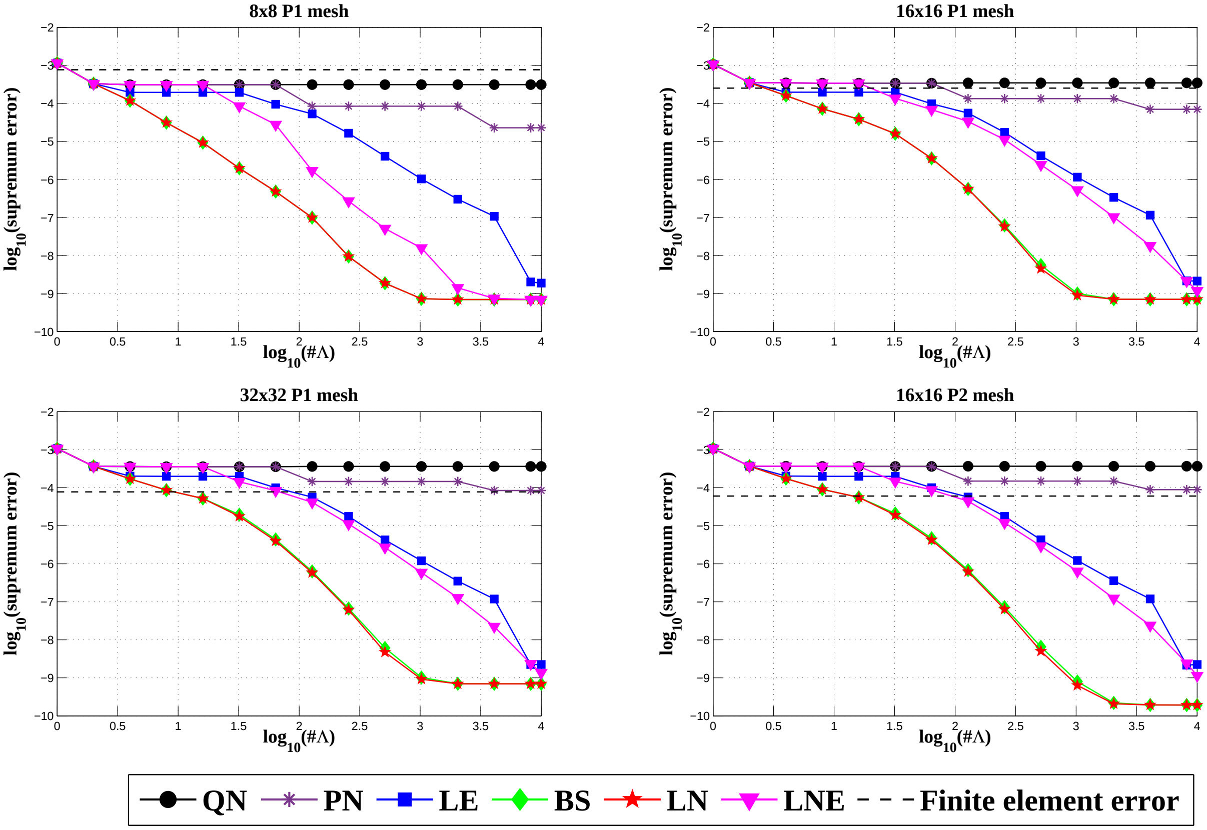


Figure 7.1: Comparison the different strategies for finite element spaces (i) (upper left), (ii) (upper right), (iii) (lower left) and (iv) (lower right).

- Second, we observe that the adaptive strategies BS and LN outperform all non adaptive strategies. They give almost identical error curves, which indicates that the LN strategy is preferable since it is has lower computational cost. In contrast, a loss in performance is observed if we instead use LNE. As to the non-adaptive strategies, LE outperforms PN and QN which do not produce any anisotropy in the coefficient sets. It is interesting to note that with 100 coefficients, the Taylor approximation error of the adaptive strategies is dominated by the finite element error, while it is still above it with 10^4 coefficients when using PN and QN.

- Finally, we observe a stagnation of order 10^{-9} in the supremum error. We interpret this by the fact that our algorithm computes once and for all the Taylor coefficients and that small numerical error resulting from linear system inversion accumulate in such computations. In turn the computed Taylor development converges towards a limit which slightly differs from $u_h(y)$.

In order to obtain a fair comparison between the different algorithms, we also show on Figure 7.2 their error curves in terms of the total number of boundary value problems which have been solved, and which is a better reflection of the CPU time (here we only consider the spatial discretization by 16×16 squares $\mathbb{P}_1$ finite elements). For non-adaptive strategies and for LNE, this number is the same as $\#(\Lambda)$, but it exceeds it moderately for LN and more strongly for BS. In this new comparison, we observe that the algorithm LN gives the best performance, followed by LNE and LE.

Remark 7.1 *Since we have observed that the error curves and selected adaptive sets do not depend much on the finite element space discretization, an interesting perspective for gaining CPU time is to first use a coarse grid finite element space to find the adaptive coefficients sets Λ_n. One may then use a finer grid for the computation of the coefficients in such sets, therefore avoiding the overhead caused by solving more boundary value problems than $\#(\Lambda_n)$ with the fine discretization. We may also use the coarse grid error curves to estimate the number of Taylor coefficients that we need to compute with the fine discretization in order to reach a prescribed accuracy.*

Remark 7.2 *Our analysis shows that we can set a stopping criterion for our adaptive algorithm based on the accuracy of the Taylor approximation to $u_h(y)$: the algorithm terminates at some step n such that*

$$\sup_{y\in U} \|u_h(y) - \sum_{\nu\in\Lambda_n} t_{\nu,h} y^\nu\|_V \leq \varepsilon,$$

where $\varepsilon > 0$ is a prescribed tolerance. A natural choice is to choose ε of the same order as the finite element error

$$\sup_{y\in U} \|u_h(y) - u(y)\|_V.$$

While this last quantity is not exactly known to us, it can be bounded by above according to a-priori estimate (6.6) *based on our knowledge of the maximal Sobolev smoothness of $u(y)$, or estimated in a finer way based on a-posteriori analysis.*

Remark 7.3 *In all six adaptive approaches, the specific choice of numbering coordinates y_j might influence the selection of the approximations once ties in certain quantities occur. In the present numerical experiments, the* 64 *coordinates were enumerated in lexicographic order according to the location of the support of the ψ_j in D. We performed the same experiments with several random reshufflings of the indexation (so that the most significant parameter y_j does not appear as first coordinate) which rendered indistinguishable results from the ones reported here; although this finding is, to some extent, implementation dependent, it strong suggests that the presented algorithms will perform well also for more general parameter dependences, where the most significant coordinate appears only in high dimension.*

In order to have an idea of the geometry of the coefficients sets Λ produced by the different strategies, we plot the projection on the two first variables, i.e. the sets

$$\{(\nu_1, \nu_2)\ ;\ \nu \in \Lambda\}.$$

Note that ν_1 and ν_2 correspond to the most "active" variables y_1 and y_2 in view of the choice of the ψ_j. We compare these sets on Figure 7.3, when $\#(\Lambda) = 200$ for the various strategies. As expected, the sets obtained for the non-adaptive choices QN and PN do not reach a high degree due to the curse of dimensionality: when $d = 64$ the dimension of the spaces $\mathbb{Q}_1$ and $\mathbb{P}_2$ clearly exceeds 200 and therefore no degree higher than 1 and 2 can be reached for any variable when using these two methods respectively. In contrast, the adaptive strategies capture the anisotropic feature of the problem and reach a high polynomial degrees in the active variables. As already mentionned, the sets obtained by BS and LN are quite similar. It is interesting to note that the geometry of the sets obtained by BS, LN and LNE significantly differs from the anisotropic simplex shape obtained with the LE strategy based on the a-priori estimates on the $\|t_\nu\|_V$.

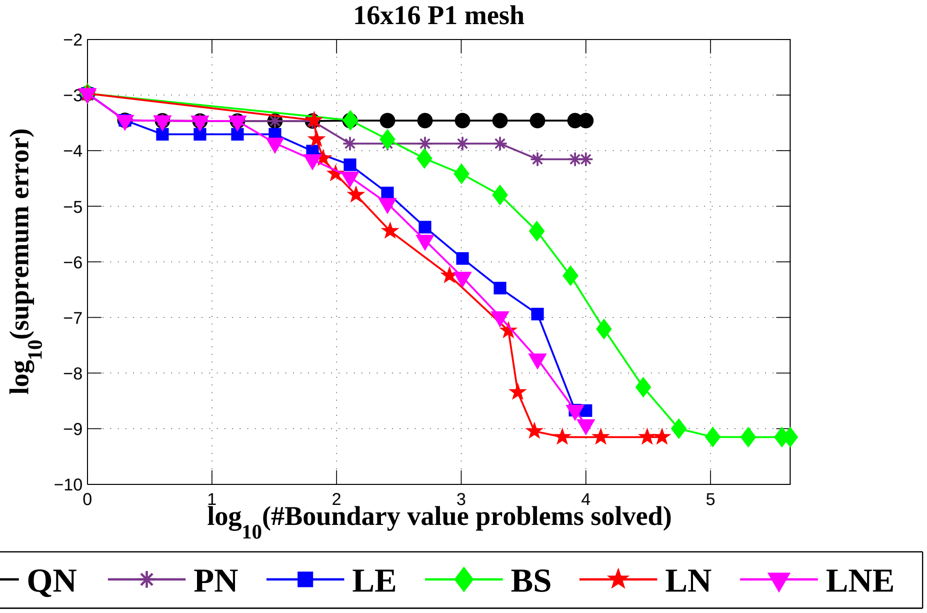


Figure 7.2: Comparison of the different strategies in term of total number of solved bvp

Finally, we have also investigated the convergence of the mean value solution $\bar{u} = \mathbb{E}(u)$ when the y_j are i.i.d. random variables which are uniformly distributed in $[-1, 1]$. Given a Taylor approximation $u_\Lambda(y) := \sum_{\nu\in\Lambda} t_\nu y^\nu$ computed for a certain set Λ by one of the proposed strategies, this mean value may thus be approximated by

$$\bar{u}_\Lambda := \sum_{\nu\in\Lambda} t_\nu \mathbb{E}(y^\nu),$$

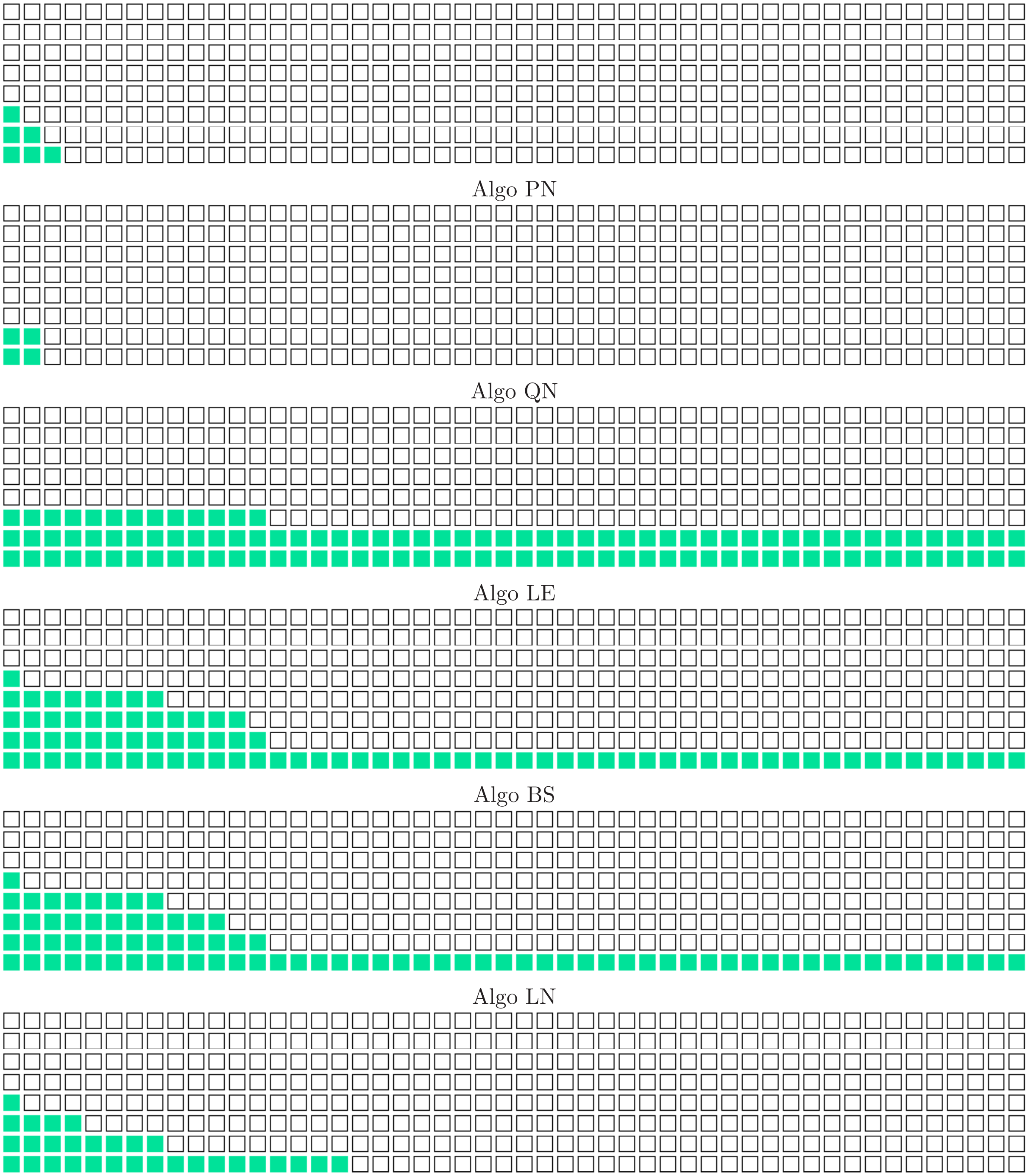
with

$$\mathbb{E}(y^\nu) = \prod_{j=1}^{d} \mathbb{E}(y_j^{\nu_j}) = \prod_{j=1}^{d} \Big(\int_{-1}^{1} t^{\nu_j} \frac{dt}{2} \Big) = \prod_{j=1}^{d} \frac{1 + (-1)^{\nu_j}}{2 + 2\nu_j} \, .$$

We are ensured that the difference between the averages $\bar{u}$ and $\bar{u}_\Lambda$ does not exceed the supremum error in y between $u(y)$ and $u_\Lambda(y)$ which was previously estimated for the various methods. Since we do not know the exact value of $\bar{u}$ for the computation of the error, we replace it by the value $\bar{u}_\Lambda$ obtained with BS algorithm when $\#(\Lambda) = 10000$, which is thus accurate up to an error of order 10^{-10}. This allows us to make the comparison between performance of the various strategies for approximating $\bar{u}$ by the error curves in terms of the number of coefficients. In addition we may compare this with the accuracy of the Monte-Carlo method, which consists in computing the empirical average

$$\bar{u}_n := \frac{1}{n} \sum_{i=1}^{n} u(y^i),$$

where $y^1, \cdots, y^n$ are independent random draws of the vector y. Since the MC method requires solving n boundary value problems, we compare its performance to the previous methods when the total number of solved boundary value problem is n, as n varies. The results are displayed on Figure 7.4. For the MC method, we display the average of the error curves for 6 independant realizations in order to illustrate the



Algo PN

Algo QN

Algo LE

Algo BS

Algo LN

Algo LNE

Figure 7.3: Comparison the index sets projected on (ν_1, ν_2) when $\#(\Lambda) = 200$ (from top to bottom: QN, PN, LE, BS, LN and LNE)

expected error $\mathbb{E}(\|\bar{u} - \bar{u}_n\|_V)$ rather than the error $\|\bar{u} - \bar{u}_n\|_V$ for a particular realization (which is more oscillatory). The $n^{-1/2}$ rate of decay of the MC method is clearly outperformed by the Taylor approximation methods based on the adaptive selection of Λ, which is rather striking in view of the large dimension $d = 64$. Note however, that in contrast to the Taylor approximation method, the MC approach allows us to solve all boundary value problems in parallel.

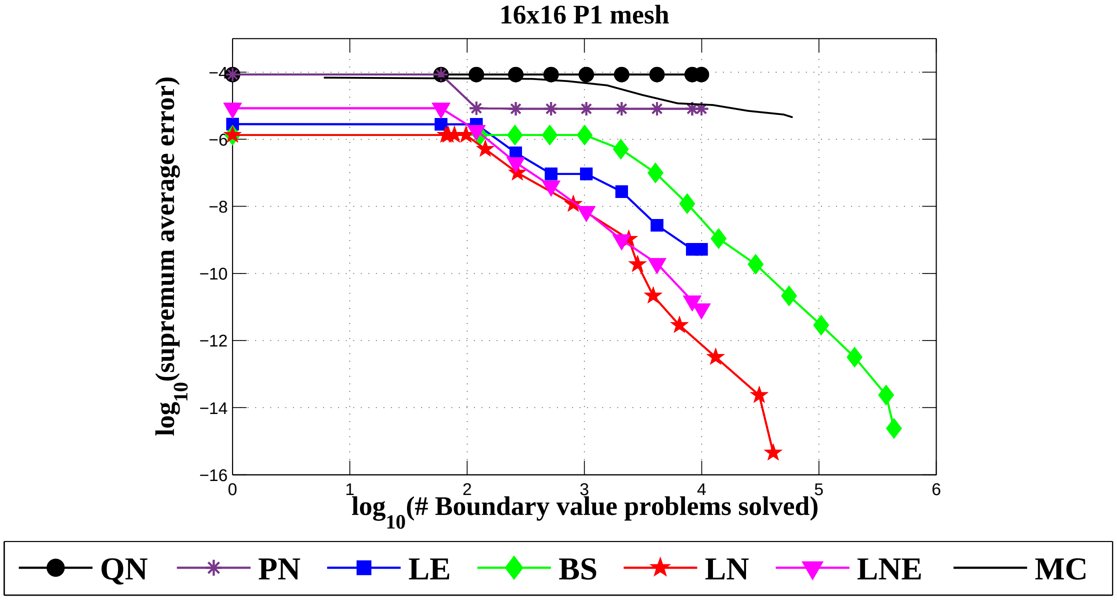


Figure 7.4: Comparison of the different strategies with Monte Carlo method.

References

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

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

[3] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Num. Math. 85-4, 579-608, 2000.

[4] M. Bieri, R. Andreev and Ch. Schwab, *Sparse Tensor Discretization of Elliptic sPDEs* SIAM J. Sci. Comput. 31, 4281-4304, 2009.

[5] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., 97, 219-268, 2004

[6] S. Brenner and L.R. Scott, *The mathematical theory of Finite Elements* (2nd Ed.), Springer, 2008.

[7] A. Buffa, Y. Mada, A. T. Patera, C. Prudhomme, and G. Turinici, *A priori convergence of the greedy algorithm for the parameterized reduced basis*, preprint, 2009.

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

[9] A. Cohen, W. Dahmen and R. DeVore, *Adaptive wavelet methods for elliptic operator equations - Convergence rates*, Math. Comp. 70, 27-75, 2000.

[10] A. Cohen, W. Dahmen and R. DeVore, *Adaptive wavelet methods for operator equations - Beyond the elliptic case*, J.FoCM 2, 203-245, 2002.

[11] A. Cohen, R. DeVore and C. Schwab, *Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs*, to appear in J. FoCM, 2010.

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

[13] R. DeVore, *Nonlinear Approximation*, Acta Numerica 7, 51-150, 1998.

[14] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal. 33, 1106-1124, 1996.

[15] Ph. Frauenfelder, Ch. Schwab and R.A. Todor: Finite elements for elliptic problems with stochastic coefficients Comp. Meth. Appl. Mech. Engg. 194, 205-228, 2005.

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

[17] T. Gantumur, H. Harbrecht and R. Stevenson, *An optimal adaptive wavelet method without coarsening of the iterands*, Math. Comp. 76, 615-629, 2007.

[18] P. Grisvard, *Elliptic problems on non-smooth domains*, Pitman, 1983.

[19] V.H. Hoang and Ch. Schwab, *Sparse tensor Galerkin discretizations for parametric and random parabolic PDEs I: Analytic regularity and gpc-approximation.* Report 2010-11, Seminar for Applied Mathematics, ETH Zürich (in review).

[20] V.H. Hoang and Ch. Schwab, *Analytic regularity and gpc approximation for parametric and random 2nd order hyperbolic PDEs*, Report 2010-19, Seminar for Applied Mathematics, ETH Zürich (to appear in Analysis and Applications (2011)).

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

[22] R. Milani, A. Quarteroni and G. Rozza, *Reduced basis methods in linear elasticity with many parameters* Comp. Meth. Appl. Mech. Engg.197, 4812-4829, 2008.

[23] P. Morin, R.H. Nochetto, and K.G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal. 38, 466-488, 2000.

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

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

[26] Ch. Schwab and A.M. Stuart *Sparse deterministic approximation of Bayesian inverse problems* , Report 2011-16, Seminar for Applied Mathematics, ETH Zürich (in review).

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

[28] R. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math. 7, 245-269, 2007.

Abdellah Chkifa
UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
chkifa@ann.jussieu.fr

Albert Cohen
UPMC Univ Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France
cohen@ann.jussieu.fr

Ronald DeVore
Department of Mathematics, Texas A& M University, College Station, TX 77843, USA
rdevore@math.tamu.edu

Christoph Schwab
Seminar for Applied Mathematics, ETH Zürich, CH 8092 Zürich, Switzerland
schwab@math.ethz.ch