Sparse polynomial methods in high dimension : application to parametric PDE

Moulay Abdellah Chkifa

Université Pierre et Marie Curie
Paris 6

Laboratoire Jacques-Louis Lions
UMR 7598

Thèse de Doctorat de
l'Université Pierre et Marie Curie

Présentée et soutenue publiquement le 14 octobre 2014

pour l'obtention du grade de
Docteur de l'Université Pierre et Marie Curie
Spécialité : Mathématiques Appliquées

par
Moulay Abdellah CHKIFA

Méthodes polynomiales parcimonieuses en grande dimension. Application aux EDP Paramétriques.

après avis des rapporteurs

M. Tony LeLievre
M. Rob Stevenson

devant le jury composé de

M.	Ivan Maday	Examinateur
M.	Christoph Schwab	Examinateur
M.	Anthony Nouy	Examinateur
M.	Albert Cohen	Directeur de Thèse
M.	Tony Lelievre	Rapporteur
M.	Rob Stevenson	Rapporteur

École Doctorale de Sciences Mathématiques de Paris Centre
ED 386

Faculté de Mathématiques
UFR 929

Moulay Abdellah CHKIFA :

UPMC, Université Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France.
CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, F-75005, Paris, France.

Adresse électronique: chkifa@ann.jussieu.fr

Remerciements

Contents

Introduction: version française

1 Equations différentielles partielles paramétriques

Cette thèse est consacrée à l'étude théorique et l'approximation numérique des équations aux dérivées partielles (EDP) paramétriques en grandes dimensions. Les EDP paramétriques apparaissent dans des contextes très variés pour la modélisation de la dépendence de systèmes physiques spécifiques en fonction de certains paramètres pertinents. Par exemple, la distibution de la chaleur dans une plaque métallique où la paramétrisation décrit le pourcentage des différentes composantes chimiques de l'alliage. La représentation formelle que nous adoptons pour les EDP paramétriques est la suivante :

$$\mathcal{D}(u, y) = 0, \tag{1.1}$$

où $\mathcal{D}$ est un opérateur différentiel linéaire ou non linéaire, modélisant le phénomène physique, qui dépend d'un ensemble de d paramètres représentés par le vecteur $y := (y_1, \dots, y_d) \in \mathbb{R}^d$. Nous désignons par $U \subset \mathbb{R}^d$ le domaine paramétrique qui décrit la plage des valeurs admissibles de y, et nous supposons qu'il existe un espace de Banach V, par exemple un espace de Sobolev, dans lequel le problème (1.1) est bien posé pour tout $y \in U$. Nous pouvons ainsi définir *l'application solution* de U dans V :

$$u : y \mapsto u(y), \tag{1.2}$$

qui associe à chaque paramètre $y \in U$ l'unique solution $u(y) \in V$ de (1.1).

Les EDP paramétriques sont utilisées dans la modélisation des systèmes complexes dans des contextes physiques ou d'ingénierie très variés. Nous ne faisons pas une classification exhaustive des ses contextes, mais nous distinguons deux classes principales:

- **Modélisation déterministe:** Les paramètres y sont des données déterministes du système physique qui peuvent être contrôlées et modifiées par l'utilisateur. Ils peuvent par exemple être des *paramètres de conception ou de contrôle* dans un processus industriel réel ou numériquement simulé. Une application typique dans ce contexte est l'optimisation d'une certaine quantité d'intérêt scalaire Q qui dépend de la solution et par conséquent des paramètres:

$$y \mapsto u(y) \mapsto Q(u(y)). \tag{1.3}$$

Par exemple, considérons l'équation stationnaire de la chaleur dans un domaine donné D

$$-\text{div}(a\nabla u) = f \quad \text{in} \quad D, \quad u_{|\partial D} = 0, \tag{1.4}$$

en présense d'une source thermique f donnée et $a = a(y)$ choisi dans une famille $\{a(y) \ : \ y \in U\}$ de fonctions de conductivité thermique. Pour des fins de conception du matériau, nous pouvons varier le paramètre y dans le but de minimiser le flux du champ de température $u(y)$ à travers une portion de la surface $\Gamma \subset \partial D$. Dans ce cas, la quantité d'intérêt scalaire est

$$y \mapsto Q(y) = \int_\Gamma \frac{\partial u(y)}{\partial n}(x)dx. \tag{1.5}$$

- **Modélisation stochastique:** Les paramètres y sont des réalisations de variables aléatoires qui reflètent des incertitudes dans le modèle physique décrit par (1.1). Par exemple, si l'équation (1.4) est utilisée pour la modélisation de la diffusion dans un milieu poreux dont les propriétés ne sont pas connues exactement, il est alors naturel de modéliser le coefficient de diffusion a comme un champ aléatoire qui, comme expliqué plus loin, peut être décrit par une suite $(y_j)_{j\geq 1}$ de variables aléatoires scalaires. Dans la modélisation stochastique, l'utilisateur est typiquement intéressé par les propriétés statistiques de la solution u, qui est elle même un champ aléatoire sur V. Par exemple, on peut être intéressé par le calcul du champ moyen $\bar{u} := \mathbb{E}[u]$ qui, si il existe, est une fonction déterministe dans V, de l'écart type $\mathbb{E}[\|u - \bar{u}\|_V^2]$, de l'espérance d'une quantité d'intérêt $Q = Q(y)$ dépendant de la solution comme dans le contexte déterministe, ou d'un intervalle de confiance pour cette quantité.

Outre la distinction entre les contextes déterministes et stochastiques, les paramètres $(y_j)_{j\geq 1}$ peuvent être utilisés dans la description de différentes quantités: la conductivité ou les propriétés de diffusion du matériau comme dans les exemples ci-dessous, le flux dans un problème de transport, un terme de forçage comme celui de droite dans (1.4), la géométrie du domaine physique (via la paramétrisation de la frontière, par exemple à l'aide des points de contrôle dans une conception assistée par ordinateur). Il est aussi possible que plusieurs de ces quantités soient simultanément considérées, auquel cas y concatène tous les paramètres utilisés dans la description de ces quantités.

Une partie importante de cette thèse est consacrée à l'étude du problème modèle (1.4) pour une classe particulière de coefficients a. Bien que simple en énoncé, ce modèle est pertinent pour la création d'une méthodologie traitant d'autres classes d'EDP paramétriques. Ici, $D \subset \mathbb{R}^m$ est un domaine lipschitzien borné, avec m typiquement égal à 2 ou à 3, et f dans $H^{-1}(D)$. Nous considérons le problème elliptique du second ordre

$$-\text{div}(a(y)\nabla u) = f \quad \text{dans} \quad D, \quad u_{|\partial D} = 0, \tag{1.6}$$

où pour tout $y \in U$, la fonction de diffusion $a(y) \in L^\infty(D)$ dépend de façon affine en y, selon

$$a(y) := \overline{a} + \sum_{j=1}^{d} y_j \psi_j \tag{1.7}$$

avec $\overline{a}$ et les ψ_j sont des fonctions dans $L^\infty(D)$. Nous faisons l'hypothèse que le problème est *uniformément elliptique* sur U, au sens où il existe $0 < r \leq R < \infty$ tels que

$$r \leq a(x,y) \leq R, \quad x \in D, \quad y \in U, \tag{1.8}$$

avec la notation

$$a(x,y) := a(y)(x) = \overline{a}(x) + \sum_{j=1}^{d} y_j \psi_j(x). \tag{1.9}$$

Sous ces hypothèses, la théorie de Lax-Milgram garantit que le problème (1.6) est bien posé dans $V := H_0^1(D)$ pour tout $y \in U$. L'application solution associe à tout $y \in U$ l'unique solution $u(y) \in V$.

Le fait de supposer une dépendance affine en y pour $a(y)$ est pertinent dans plusieurs contextes. Par exemple si a est constant par morceaux sur une partition $D = \cup_{j=1}^{d} D_j$ du domaine physique, alors il est naturel de poser

$$a(y) = \overline{a} + \sum_{j=1}^{d} y_j \chi_{D_j}, \tag{1.10}$$

où $\overline{a}$ est une constante et χ_{D_j} la fonction indicatrice de D_j. Plus généralement, la forme affine (1.7) est rencontrée si nous tronquons le développement de $a - \overline{a}$, où $\overline{a}$ est une fonction de x, dans une base donnée $(\psi_j)_{j\geq 1}$, c'est à dire un développement de la forme

$$a(y) = \overline{a} + \sum_{j=1}^{\infty} y_j \psi_j. \tag{1.11}$$

Il existe évidement plusieurs choix possibles pour une telle base (séries de Fourier, polynômes orthogonaux, ondelettes...). Dans le contexte stochastique, lorsque a est un champ aléatoire du second ordre avec espérance $\mathbb{E}[a] = \overline{a}$ et avec une fonction de covariance continue

$$(x,z) \in D \times D \mapsto cov[a](x,z) := \mathbb{E}[(a(x) - \overline{a}(x))(a(z) - \overline{a}(z))], \tag{1.12}$$

un choix fréquemment utilisé est la *base de Karhunen-Loève*, en d'autres termes, les vecteurs propres orthonormaux de l'opérateur

$$v \mapsto T_a v := \int_D cov[a](\cdot, x) v(x) dx, \qquad v \in L^2(D), \tag{1.13}$$

qui est compact, auto-adjoint et positif sur $L^2(D)$. Les variables scalaires y_j sont centrées et mutuellement non-corrélées, i.e. $\mathbb{E}[y_j] = 0$ et $\mathbb{E}[y_i y_j] = \delta_{ij}$ pour $i, j \geq 1$, avec variance donnée par la valeur propre correspondante $\lambda_j > 0$.

Tout au long de cette thèse, nous supposons que le domaine paramétrique U a une forme tensorielle, ce qui veut dire que les variables y_j varient indépendamment dans des intervalles I_j. Une telle hypothèse est naturelle pour des problèmes déterministes où ces paramètres peuvent être indépendamment ajustés. Par exemple, dans le cas constant par morceaux (1.10), ces intervalles peuvent être de la forme $I_j = [-\alpha_j, \alpha_j]$ avec $0 < \alpha_j < \overline{a}$ pour tout j. Dans le contexte stochastique, en utilisant par exemple la représentation Karhunen-Loève ci-dessous, cette supposition est naturelle si on suppose que les y_j sont des variables aléatoires indépendantes. Notons que l'indépendance statistique des composantes est une propriété plus forte que leur non-corrélation. Nous pouvons donc, quitte à changer la normalisation des fonctions de base ψ_j, supposer dans les deux contextes, déterministe et stochastique, que le domaine des paramètres U est l'hypercube unitaire en dimension d,

$$U := [-1, 1]^d. \tag{1.14}$$

Dans les modèles où les paramètres $y = (y_1, \ldots, y_d)$ correspondent à la troncation d'une series infinie telle que (1.11), la précision est affectée par l'ordre d de la troncation. Afin d'atteindre des précisions arbitrairement élevées dans l'approximation numérique de ces modèles, il faut par conséquent autoriser la croissance de la variable d. Comme expliqué plus loin, cette croissance a en principe un coût numérique sévère exprimé par *la plaie des grandes dimensions*. Un des objectifs de cette thèse est donc de développer des méthodes numériques qui sont robustes à la croissance de d, au sens où elles peuvent être appliquées au cas où le vecteur paramètre

$$y = (y_j)_{j \geq 1}, \tag{1.15}$$

est de dimension infinie. Dans ce cas, le domaine paramétrique est l'hypercube en dimension infinie,

$$U := [-1, 1]^{\mathbb{N}}. \tag{1.16}$$

2 Approximation numérique

Dans le contexte déterministe comme stochastique, les applications concrètes peuvent exiger en principe l'évaluation de la solution $u(y)$ pour un grand nombre N d'instances du paramètre vectoriel y. Les exemples typiques dans ce sens sont l'optimisation d'une quantité d'intérêt scalaire $y \mapsto Q(u(y))$ par la méthode de Newton, où l'approximation de la moyenne $\mathbb{E}[Q(u(y))]$ par les méthodes de Monte Carlo. De telles approches exigent des résolutions $\{u_i = u(y^i) \ : \ i = 1, \ldots, N\}$ de l'application solution (1.2), chacune

d'entre elle étant effectuée par un solveur numérique, éventuellement très coûteux dans le cas d'un système complexe. Ajoutons à cela que les approches citées sont souvent *orientées objectif*, en d'autres termes, la base de données des évaluations collectées pour une certaine tâche (par exemple l'optimisation) peut être mal adaptée pour une autre tâche (par exemple le calcul de la moyenne).

Les difficultés décrites ci-dessus conduisent au défi d'approcher simultanément toutes les solutions $u(y)$ pour $y \in U$ à une précision prescrite et avec un coût numérique raisonnable, ce qui revient à approcher l'application solution $u : y \mapsto u(y)$.

Cette tâche est difficile car, contrairement au problème de l'approximation d'une fonction à valeurs réelles $u : \mathbb{R} \mapsto \mathbb{R}$, l'application solution u associée à une EDP paramétrique (i) est définie sur un domaine multi-dimensionnel $[-1, 1]^d$ où la dimension paramétrique d peut être large ou même infinie, et (ii) prend ses valeurs dans un espace V de dimension infinie, ou dans un espace de discrétisation $V_h \subset V$ de dimension finie mais grande quand un solveur numérique est utilisé.

Le premier point (i) souligne le problème de la plaie des grandes dimensions qui fait référence à l'explosion exponentielle de la complexité des méthodes de discrétisation, avec la croissance du nombre d de variables, même pour les fonctions à valeurs réelles. Une autre émanation de ce phénomène est la détérioration des taux d'approximation des fonctions d'une régularité donnée quand d croît: par exemple la précision en métrique L^∞ (ou uniforme) de la reconstruction d'une fonction arbitraire avec des dérivées continues jusqu'à l'ordre m avec des polynômes par morceaux dans des grilles h-espacées est au mieux de l'ordre de h^m et donc, en terme du nombre de degrés de liberté n, est d'ordre asymptotique $n^{-m/d}$. La vitesse de convergence est donc d'autant plus médiocre que d est grand. Un examen plus approfondi à l'aide de la théorie des épaisseurs non-linéaires [43, 40, 80] révèle que ce taux de convergence ne peut pas être amélioré avec une autre méthode de discrétisation.

Le deuxième point (ii) est lié au calcul pratique des approximations. Les instances $u(y)$ de l'application solution ou toute quantité qui en dépend, par exemple les coefficients d'une approximation polynomiale en la variable paramétrique y de cette application, peuvent seulement être approchées avec une certaine discrétisation de l'espace, telle que par la méthode des éléments finis. Par conséquent, il est crucial d'incorporer ces considérations dans l'analyse de l'erreur numérique finale. Plusieurs questions peuvent être soulevées dans l'analyse de l'erreur de discrétisation. Par exemple, est-il judicieux d'utiliser le même espace de discrétisation V_h pour toutes les instances? est ce que la méthode d'approximation de l'application solution u est robuste aux erreurs de discrétisation? etc. Nous laissons de côté la discrétisation de l'espace dans le reste de l'introduction et nous nous concentrons sur la discrétisation paramétrique (en y).

Nous distinguons deux approches dans l'approximation de l'application solution (1.2). Une propriété commune à ces deux approches est la séparation du paramètre vectoriel y et de la variable physique x, espace et/ou temps, dans l'approximation de

u. La première approche consiste à construire une application peu coûteuse à évaluer

$$y \in U \mapsto u_n(y) := \sum_{i=1}^{n} v_i \phi_i(y) \in V, \tag{2.1}$$

basée sur un petit nombre n de fonctions $v_i \in V$ et de fonctions à valeurs scalaires ϕ_i de U dans $\mathbb{R}$ ou $\mathbb{C}$. Par exemple, les v_i peuvent être des instances de l'application solution u associées avec des valeurs $y^j \in U$ bien choisies du paramètre vectoriel, i.e. $v_i = u(y^i)$, et les fonctions ϕ_i dans ce cas sont des fonctions de Lagrange associées avec un schéma d'interpolation par des polynômes aux points $(y^1, \ldots, y^n)$. Dans le contexte stochastique, ces méthodes sont communément appelées *méthodes stochastiques spectrales*, voir [51, 52, 59].

En fonction du contexte de modélisation, déterministe ou stochastique, et de l'application visée, on décide en quel sens l'approximation u_n doit être proche de u. Si par exemple l'objectif est de capturer l'application u partout sur U à une précision prescrite $\varepsilon(n)$, alors l'erreur doit être considérée dans le sens *uniforme*, i.e.

$$\sup_{y \in U} \|u(y) - u_n(y)\|_V \leq \varepsilon(n). \tag{2.2}$$

Dans le contexte stochastique, la qualité de l'approximation est souvent mesurée en *moyenne*, par exemple à travers une estimation de l'erreur quadratique moyenne de la forme

$$\mathbb{E}[\|u(y) - u_n(y)\|_V^2] := \int_U \|u(y) - u_n(y)\|_V^2 d\varrho(y) \leq \varepsilon^2(n), \tag{2.3}$$

où ϱ est la distribution de probabilité du vecteur aléatoire y. Notons que la première estimation entraîne la seconde.

La deuxième approche consiste à rechercher un sous espace E_n de V de dimension faible n qui peut servir pour l'approximation simultanée de toutes les solutions $u(y)$, par exemple par la méthode de Galerkin. Ceci signifie que nous cherchons à approcher la *variété des solutions*,

$$\mathcal{M} := \left\{ u(y) : y \in [-1, 1]^d \right\} \subset V, \tag{2.4}$$

par l'espace vectoriel E_n. À nouveau, on peut chercher les estimations d'erreur au sens uniforme ou dans un sens quadratique moyen, entre u et sa meilleure approximation $u_{E_n} : y \mapsto u_n(y) := \operatorname{argmin}_{v \in E_n} \|u(y) - v\|_V$, obtenue par la projection orthogonale de chaque $u(y)$ dans E_n dans le cas où V est un espace de Hilbert.

Dans le cas de l'approximation au sens uniforme, le choix optimal de E_n lorsqu'il existe correspond à l'espace qui réalise la *n-épaisseur de Kolmogorov*,

$$d_n(\mathcal{M})_V := \inf_{\dim(E) \leq n} \sigma_E(\mathcal{M}), \qquad \sigma_E(\mathcal{M}) := \sup_{w \in \mathcal{M}} \inf_{v \in E} \|w - v\|_V. \tag{2.5}$$

Dans le contexte stochastique, on soustrait habituellement le champ moyen $\overline{u} = \mathbb{E}[u]$ à u et on cherche l'espace E_n qui minimise l'erreur quadratique moyenne

$$\mathbb{E}(\|\tilde{u} - \tilde{u}_E\|_V^2), \tag{2.6}$$

entre $\tilde{u} = u - \overline{u}$ et sa meilleure approximation $\tilde{u}_E$, parmis tous les sous-espaces E de dimension n. Le choix optimal est lié au développement de Hilbert-Karhunen-Loève

$$\tilde{u} = \sum_{j=1}^{\infty} \sqrt{\lambda_i} v_j U_j \tag{2.7}$$

où (λ_i, v_i) est la famille des couples, de valeurs propres classées par ordre décroissant et de vecteurs propres associés à l'opérateur de covariance de u sur V (voir [45] pour plus de détails), et

$$U_j := \frac{1}{\sqrt{\lambda_i}} \langle \tilde{u}, v_j \rangle_V, \tag{2.8}$$

sont centrées et mutuellement non-corrélées, et de variance 1. L'espace optimal est alors engendré par $\{v_1, \ldots, v_n\}$.

Dans les deux cas, les espaces optimaux ne sont pas facilement accessibles d'un point de vue numérique et on a donc recours à des espaces sous-optimaux mais facile à calculer. Pour des estimations au sens quadratique moyen, on peut approcher $\overline{u}$ et le noyau de covariance $cov[u]$ à partir de la connaissance de u sur une discrétisation grossière de V, voir [45], où par le calcul de $\overline{u}$ et le noyau de covariance $cov[u]$ en n'ayant aucune connaissance de u, voir [65] pour le problème (1.6). Cependant, le calcul du développement Hilbert-Karhunen-Loève revient à la résolution d'un problème aux valeurs propres généralisé, ce qui peut être coûteux numériquement. Pour les estimations uniformes, une stratégie populaire est la méthode des *bases réduites*, [16, 64, 63]. Dans cette stratégie, on calcule en premier lieu au cours d'une étape *offline*, éventuellement très coûteuse, n éléments $v_j = u(y^j)$ de la variété $\mathcal{M}$, puis dans le calcul *online* on approche pour n'importe quel paramètre y la solution $u(y)$ dans l'espace $F_n = \text{span}\{v_1, \ldots, v_n\}$ par un schéma de Galerkin. Les espaces F_n sont en général sous-optimaux par comparaison avec les espaces des n-épaisseur E_n qui réalisent l'infimum en (2.5). Cependant, il a été prouvé dans [12] et [42] qu'une certaine sélection *greedy* des instances v_j dans l'étape offline produit des espaces qui sont optimaux en taux de convergence, au sens suivant: étant donnés $\alpha, c > 0$, il est prouvé que

$$\sup_{n>0} n^\alpha \sigma_{F_n}(\mathcal{M}) \leq C_\alpha \sup_{n>0} n^\alpha d_n(\mathcal{M}), \tag{2.9}$$

et que

$$\sup_{n>0} e^{c_\alpha c n^\alpha} \sigma_{F_n}(\mathcal{M}) \leq C_\alpha \sup_{n>0} e^{c n^\alpha} d_n(\mathcal{M}), \tag{2.10}$$

où C_α et c_α sont des constantes qui dépendent de α.

Il faut remarquer que les deux approches décrites ci-dessus sont liées. D'une part, si u_n est une approximation de u de la forme (2.1) dans le sens uniforme à une précision $\varepsilon(n)$, alors en introduisant l'espace $F_n := \text{span}\{v_1, \ldots, v_n\}$, nous avons

$$\sigma_{F_n}(\mathcal{M}) = \sup_{y\in U} \inf_{v\in F_n} \|u(y) - v\|_V \leq \sup_{y\in U} \|u(y) - u_n(y)\|_V \leq \varepsilon(n). \tag{2.11}$$

Par conséquent, l'espace vectoriel F_n approche la variété $\mathcal{M}$ dans le sens uniforme avec la précision $\varepsilon(n)$, entraînant ainsi une estimation de la n-épaisseur de Kolgomorov par

$$d_n(\mathcal{M}) \leq \sigma_{F_n}(\mathcal{M}) \leq \varepsilon(n). \tag{2.12}$$

Ceci peut à son tour être utilisé pour l'étude de la convergence de la méthode des bases réduites, au vu de (2.9) et (2.10). D'autre part, si $E_n := \text{span}\{v_1, \ldots, v_n\}$ est l'espace qui réalise la n-épaisseur de Kolmogorov, lorsqu'elle atteinte, alors on peut écrire pour tout $y \in U$

$$u_{E_n}(y) = \sum_{i=1}^{n} v_i \phi_i(y), \tag{2.13}$$

où $\phi_i(y)$ sont les coordonnées de $u_{E_n}(y)$ associées avec les v_i. Evidemment u_{E_n} a la forme (2.1) et

$$d_n(\mathcal{M}) = \sigma_{E_n}(\mathcal{M}) = \sup_{y\in U} \|u(y) - u_{E_n}(y)\|_V, \tag{2.14}$$

ce qui montre que u_{E_n} est la meilleure approximation de la forme (2.1) dans le sens uniforme.

En pratique, il y a cependant une différence fondamentale entre les deux approches. Dans la première approche, on choisit d'abord n fonctions à valeurs scalaires $\phi_1, \ldots, \phi_n$ dans une famille de fonctions, par example la famille des polynôme de Legendre, puis on calcule les fonctions $v_1, \ldots, v_n \in V$. Dans la deuxième approche, on s'applique à identifier n "bonnes" fonctions $v_1, \ldots, v_n$ dans V, puis pour y donné on calcule les valeurs $\phi_j(y)$ par une méthode de projection de Galerkin dans l'espace engendré par ces fonctions. Une fois les fonctions $v_1, \ldots, v_n \in V$ sont calculées, la première approche présente l'avantage du fait que le calcul de $u_n(y)$ est immédiat par la combinaison linéaire (2.1) alors qu'une inversion de système est requise, pour tout $y \in U$, dans la deuxième approche.

Cette thèse aborde uniquement la première approche, et plus particulièrement la construction d'approximations peu coûteuses de la forme (2.1) à l'application solution, avec les fonctions $\phi_i(y)$ qui sont des polynômes multidimensionels en la variable y. Une autre distinction essentielle dans les méthodes numériques pour les EDP paramétriques fait apparaître les deux classes suivantes considérées dans cette thèse:

- ***Méthodes non-intrusives***: elles demandent des résolutions (approchées) répétées de l'application solution u, obtenues par un solveur numérique déterministe existant.

Typiquement, le solveur est un code de simulation industriel considéré comme une boite noire qui associe à chaque paramètre vectoriel $y \in U$ une evaluation $u^\varepsilon(y) \in V$ approchant $u(y)$ à une précision $\varepsilon > 0$ arbitraire requise. Un exemple de ces méthodes est l'approximation Monte Carlo du champ moyen $\overline{u}$ par moyenne empirique

$$\overline{u}_n := \frac{1}{n}\sum_{i=1}^{n} u(y^i), \tag{2.15}$$

où les y^i sont des réalisations indépendantes du paramètre vectoriel aléatoire y. Un autre exemple est l'approximation de l'application solution par l'interpolation en des points $y^1, \ldots, y^n \in U$. Les méthodes non-intrusive présentent l'intérêt de pouvoir être utilisées comme un *post-traitement* aux solveurs numériques existants. Cependant, la dépendance au solveur, qui est éventuellement coûteux numériquement, peut s'avérer contraignante. De plus, lorsque le solveur est donné comme une boite noire sans information précise sur le modèle EDP sous-jacent (1.1), il n'y a aucune garantie théorique qu'une méthode non-intrusive convient à l'objectif numérique ou même que les approximations produites par une telle méthode convergent.

- ***Méthodes intrusives***: à l'inverse des méthodes non-intrusives, elles exploitent les particularités des classes spécifiques d'EDP paramétriques ou stochastiques. Par "particularités", on entend ici toute information additionnelle (donnée ou hypothèse) sur l'équation (1.1) qui gouverne le système physique. Par exemple, la distribution du vector aléatoire y dans le contexte stochastique, la catégorie de l'opérateur $\mathcal{D}$, la forme de sa dépendance en y, etc. La connaissance de telles spécificités permet l'élaboration de méthodes qui sont bien adaptées au problème et qui peuvent être plus performantes que les méthodes non-intrusives en vitesse et en précision. Par exemple, la connaissance du modèle exacte permet l'utilisation de la méthode de Galerkin avec les polynômes en y pour la discrétisation paramétrique. Notons que la méthode des bases réduites est non-intrusive dans l'étape offline où on calcule seulement des instances de solutions, alors que l'étape online est intrusive car on utilise ces instances pour générer un espace de discrétisation pour la méthode de Galerkin.

3 Approximations polynomiales de l'application solution

Nous avons vu qu'il existe des approximations à n termes qui ont une forme séparable (2.1) et qui sont optimales au sens uniforme ou au sens quadratique moyen. En revanche, elles ne sont pas facilement calculables. En outre, les fonctions ϕ_j dans ces cas peuvent être assez complexes puisqu'elles dépendent de toutes les variables y_j. Notre approche consiste à chercher davantage de séparation de variables, via des approximations de la

forme

$$u_n(y) = \sum_{j=1}^{n} v_j \prod_{i\geq 1} \phi_{j,i}(y_i). \tag{3.1}$$

Le problème du choix optimal des facteurs est mal posé. À titre d'illustration, nous considérons le cas où u est à valeurs réelles et que le domaine paramétrique U est discret et fini de la forme $\{t_1, \ldots, t_k\}^d$. Dans ce cas $V := \mathbb{R}$ et $u \in V^{k\times\cdots\times k}$ est un tenseur d'ordre d, ainsi les approximations de la forme (3.1) sont les approximations de rang n de u qui s'écrivent sous la forme

$$u_n = \sum_{j=1}^{n} v_j \ \Phi_{j,1} \otimes \ldots \otimes \Phi_{j,d}, \tag{3.2}$$

où $\Phi_{j,i} := (\phi_{j,i}(t_1), \ldots, \phi_{j,i}(t_k))^t$. Il est bien connu que le problème du calcul de la meilleure approximation de rang $n \geq 2$ est en général mal posé pour toutes les normes, excepté dans le cas $d = 2$ où, grâce au théorème d'Eckart-Young, voir [38], il est complètement résolu pour les normes de Hilbert-Schmidt et les normes spectrales.

Les approches alternatives consistent à choisir les fonctions $\phi_{j,i}$ dans une famille prédéfinie. Par exemple, nous pouvons imposer aux fonctions $\phi_{j,i}$ d'être des polynômes, des polynômes par morceaux, des fonctions trigonométriques, etc. Ici, nous considérons essentiellement les polynômes, auquel cas les approximations de la forme (3.1) sont des polynômes à plusieurs variables définis sur U et à coefficients dans V.

Les approximations par des polynômes tensorisés peuvent être construites à l'aide de plusieurs méthodes, par exemple les séries de Taylor, les séries de Legendre, la projection de Galerkin, l'interpolation polynomiale, les moindres carrés, les grilles parcimonieuses (sparse grids), etc. Le choix d'une méthode se fait en fonction du but recherché dans l'approximation et des informations dont on dispose sur l'EDP paramétrique. En particulier, parmi toutes les méthodes que nous venons de citer, seules les trois dernières sont non-intrusives. Pour chaque méthode, les polynômes $\phi_{j,i}$ ont une forme prédéfinie. Nous introduisons dans ce qui suit une notation unifiée que nous utiliserons pour décrire les différentes méthodes polynomiales qui ont été proposées depuis quelques années.

Nous considérons une famille de polynômes à une variable

$$\mathcal{P} := (P_j)_{j\geq 0}, \tag{3.3}$$

avec P_0 constant et égale à 1 et P_j est de degré j, de sorte que $\{P_0, \ldots, P_k\}$ est une base de $\mathbb{P}_k$ l'espace des polynômes de degré au plus k. Nous désignons par $\mathcal{F}$ l'ensemble $\mathbb{N}^d$ des multi-indices de longueur d. Nous considérons la famille $(P_\nu)_{\nu\in\mathcal{F}}$ des polynômes à plusieurs variables définis par

$$P_\nu(y) := \prod_{j=1}^{d} P_{\nu_j}(y_j), \quad y := (y_1, \ldots, y_d), \qquad \nu \in \mathcal{F}. \tag{3.4}$$

Les approximations polynomiales que nous considérons ont la forme

$$u_\Lambda := \sum_{\nu\in\Lambda} v_\nu \, P_\nu, \tag{3.5}$$

où $\Lambda \subset \mathcal{F}$ est un ensemble de multi-indices de cardinalité finie et $\{v_\nu\}_{\nu\in\Lambda}$ sont des éléments de V. Les polynômes u_Λ appartiennent ainsi à l'espace des polynômes à valeurs dans V

$$\mathbb{V}_\Lambda := V \otimes \mathbb{P}_\Lambda \quad \text{où} \quad \mathbb{P}_\Lambda = \mathbb{P}_\Lambda(\mathcal{P}) := \text{span}\{P_\nu : \nu \in \Lambda\}. \tag{3.6}$$

Il est important de signaler que l'espace $\mathbb{V}_\Lambda$ dépend à priori à la fois de l'ensemble Λ et de la famille $\mathcal{P}$. Pour la famille $\mathcal{P}$, nous pouvons typiquement considérer les monômes

$$P_j(t) = t^j, \tag{3.7}$$

où les polynômes de Legendre obtenus par l'orthogonalisation Gram-Schmidt de ses derniers dans $L^2([-1,1], \frac{dt}{2})$. En revanche, si Λ a la propriété suivante:

$$\nu \in \Lambda \;\text{ et }\; \mu \leq \nu \Rightarrow \mu \in \Lambda, \tag{3.8}$$

où $\mu \leq \nu$ signifie que $\mu_i \leq \nu_i$ pour tout $i = 1, \ldots, d$, alors il facile à voir que $\mathbb{V}_\Lambda$ est indépendant de $\mathcal{P}$, auquel cas on a

$$\mathbb{V}_\Lambda := V \otimes \text{span}\{y \mapsto y^\nu : \nu \in \Lambda\} \quad \text{où} \quad y^\nu = \prod_{j=1}^{d} y_j^{\nu_j}. \tag{3.9}$$

Les ensembles Λ qui ont la propriété précédente sont appelés *ensembles bas* ou *ensembles fermés vers le bas* (lower sets ou downward closed sets), et jouent un rôle très important dans cette thèse.

Étant donné une EDP paramétrique ou stochastique (1.1), deux questions fondamentales sont à poser:

(i) Comment trouver Λ de faible cardinalité tel que l'application solution est bien approchée dans $\mathbb{V}_\Lambda$?

(ii) Comment calculer en pratique une approximation de l'application solution dans $\mathbb{V}_\Lambda$?

Dans le cas du problème elliptique modèle (1.6) avec une dépendance affine comme dans (1.7) et une hypothèse d'ellipticité uniforme (1.8), ces questions ont été abordées depuis quelques années par diverses approches [2, 5, 6, 4, 68, 70, 69, 7, 8, 74, 34, 33, 22]. Les algorithmes réalisables de l'approximation polynomiale considérés dans ces travaux intègrent la discrétisation en la variable d'espace, prenant ainsi en compte le fait que les instances $u(y) \in V$ ou tout autre coefficient dans V associé à u ne peuvent qu'être approchées avec une précision donnée. Par exemple, pour le modèle (1.6), ceci peut

passer par une discrétisation utilisant la méthode des élément finis (FEM), où les fonctions $v \in V = H_0^1(D)$ sont approchées dans des espaces $V_h \subset V$ de fonctions linéaires continues par morceaux sur un maillage régulier quasi-uniforme de D de taille de pas h (voir par exemple [28] pour une introduction générale). L'erreur FEM est ensuite prise en compte dans l'analyse de l'erreur totale de l'approximation. À ce stade, ce point n'est pas central dans notre présentation, c'est pourquoi nous décrivons les méthodes polynomiales dans un cadre semi-discret, c'est-à-dire en considérant seulement la discrétisation dans la variable paramétrique.

L'approche proposée dans [2] est basée sur le développement en séries de Neumann appliqué à la formulation suivante du problème:

$$A(y)u = f, \tag{3.10}$$

où pout tout $y \in U$, $A(y)$ est l'opérateur différentiel de V dans V^* défini par $A(y)v = -\mathrm{div}(a(y)\nabla v)$. Au vu de (1.7), cet opérateur peut être décomposé en $A = A_0 + \Psi$ où A_0 donné par $A_0 v = -\mathrm{div}(\overline{a}\nabla v)$ ne dépend pas de y et $\Psi(y)v = -\mathrm{div}((a(y) - \overline{a})\nabla v)$. Sous l'hypothèse d'ellipticité uniforme (1.8), nous pouvons écrire

$$u(y) = A(y)^{-1} f = (Id + A_0^{-1}\Psi(y))^{-1} g, \qquad g := A_0^{-1} f, \tag{3.11}$$

où $\|A_0^{-1}\Psi\|_{V\to V} \leq \xi = 1 - \frac{r}{R} < 1$. Ceci permet d'appliquer le développement en séries de Neumann et d'obtenir la borne exponentielle

$$\sup_{y\in U} \left\| u(y) - \sum_{j=0}^{k} (-1)^j (A_0^{-1}\Psi(y))^j g \right\|_V \lesssim \xi^k. \tag{3.12}$$

Puisque l'opérateur $\Psi(y)$ dépend linéairement de y, alors le polynôme dans l'approximation ci-dessus appartient à $\mathbb{V}_{\mathcal{S}_k}$ où $\mathcal{S}_k$ est le simplexe

$$\mathcal{S}_k := \left\{ \nu \in \mathbb{N}^d : |\nu| := \sum_{j=1}^{d} \nu_j \leq k \right\}. \tag{3.13}$$

On note que $\mathcal{S}_k$ est un ensemble fermé vers le bas et que $\mathbb{P}_{\mathcal{S}_k}$ est l'espace des polynômes à d variables de degré totale k. L'approximation polynomiale converge avec une vitesse exponentielle vers u dans le sens uniforme. En revanche, l'espace de polynômes est de dimension $\binom{k+d}{k}$ qui croît par conséquent rapidement avec d est k.

Dans les travaux ultérieurs [5, 6], les approximations de u dans le sens quadratique moyen sont construites par la projection de Galerkin dans des espaces prédéfinis de polynômes par morceaux et de polynômes. Dans le contexte stochastique, si on désigne par ϱ la distribution de probabilité jointe du vecteur aléatoire y, l'application u peut être définie comme l'unique fonction de l'espace de Bochner $\mathcal{V}_2 := L^2(U, V, d\varrho)$, solution du problème variationnel

$$\int_U \int_D a(y)\nabla u(y)\nabla w(y) d\varrho(y) = \int_U \int_D f w(y) d\varrho(y), \qquad w \in \mathcal{V}_2, \tag{3.14}$$

et son approximation Galerkin dans $\mathbb{V}_\Lambda = V \otimes \mathbb{P}_\Lambda$ est l'unique $u_\Lambda \in \mathbb{V}_\Lambda$ tel que

$$\int_U \int_D a(y)\nabla u_\Lambda(y)\nabla w(y)d\varrho(y) = \int_U \int_D fw(y)d\varrho(y), \qquad w \in \mathbb{V}_\Lambda. \tag{3.15}$$

Les espaces de polynômes considérés dans [5, 6] sont de type $\mathbb{V}_{\mathcal{B}_\mu}$ où pour $\mu \in \mathbb{N}^d$, $\mathcal{B}_\mu$ est le bloc rectangulaire

$$\mathcal{B}_\mu := \Big\{\nu \in \mathbb{N}^d : \nu \leq \mu\Big\}. \tag{3.16}$$

On note que $\mathcal{B}_\mu$ est fermé vers le bas et que $\mathbb{P}_{\mathcal{B}_\mu}$ est l'espace de polynômes à d variables de degré au plus μ_j dans la variable y_j. Les auteurs supposent que ϱ est égale au produit de la mesure uniforme $\hat{\varrho} := \otimes_{j=1}^d \frac{dy_j}{2}$ avec une fonction bornée. Ceci, combiné avec l'optimalité de la projection de Galerkin $u_{\mathcal{B}_\mu}$ et l'hypothèse d'ellipticité uniforme, implique

$$\mathbb{E}\Big[\|u(y) - u_{\mathcal{B}_\mu}(y)\|_V^2\Big] \lesssim \int_U \|u(y) - \sum_{\nu\in\mathcal{B}_\mu} v_\nu L_\nu(y)\|_V^2 d\hat{\varrho}(y) = \sum_{\nu\notin\mathcal{B}_\mu} \|v_\nu\|_V^2, \tag{3.17}$$

où $(L_\nu)_{\nu\in\mathcal{F}}$ sont les polynômes de Legendre tensorisés, orthonormales par rapport à $\hat{\rho}$, et les v_ν sont les coefficients associés. En obtenant des estimés sur les quantités $\|v_\nu\|_V$ via l'étude des dérivées partielles de u et en utilisant la structure produit des espaces de polynômes $\mathbb{P}_{\mathcal{B}_\mu}$, les auteurs montrent que $u_{\mathcal{B}_\mu}$ converge vers u avec une borne de la forme

$$\mathbb{E}\Big[\|u(y) - u_{\mathcal{B}_\mu}(y)\|_V^2\Big] \lesssim \sum_{j=1}^d \Big(1 + \frac{c}{\|\psi_j\|_{L^\infty(D)}}\Big)^{-(\mu_j+1)}, \tag{3.18}$$

où c est une constante fixée. Notons par contre que si $\mu_j \geq 1$ pour tout j, alors la dimension de l'espace des polynômes $\mathbb{P}_{\mathcal{B}_\mu}$ dépasse 2^d, ce qui reflète la plaie des grandes dimensions.

Dans le cas où ϱ est une mesure produit, ce qui est équivalent à l'indépendance des variables aléatoires y_j, les auteurs proposent d'utiliser des *polynômes doublement orthogonaux* tensorisés dans le but de découpler le système de Galerkin et calculer rapidement la projection de Galerkin. Sans perte de généralité, supposons que $\varrho := \otimes_{j=1}^d \frac{dy_j}{2}$ et désignons par $(L_j)_{j\geq 1}$ les polynômes de Legendre orthonormés dans $L^2([-1,1], \frac{dt}{2})$. Puisque $\mathbb{P}_{\mathcal{B}_\mu}$ a une structure produit, alors

$$\mathbb{P}_{\mathcal{B}_\mu} = \otimes_{j=1}^d \mathrm{span}\Big\{l_k^{\mu_j+1} \ : \ k = 0, \ldots, \mu_j\Big\} = \mathrm{span}\{l_\nu^\mu : \nu \leq \mu\} \tag{3.19}$$

où

$$l_k^n := \frac{L_n}{(t - t_k^n)L_n'(t_k^n)} \quad \text{et} \quad l_\nu^\mu := \otimes_{j=1}^d l_{\nu_j}^{\mu^j+1}, \qquad \nu \leq \mu, \tag{3.20}$$

et pour tout $n \geq 1$, $t_0^n, \ldots, t_{n-1}^n$ sont les n racines simples du polynôme de Legendre L_n. Il est facile de vérifier, par des arguments élémentaires d'orthogonalité que pour tout $n \geq 1$ et $0 \leq i, j \leq n-1$,

$$\int_{-1}^{1} l_i^n(t) l_j^n(t) \frac{dt}{2} = \beta_i^n \delta_{i,j} \quad \text{et} \quad \int_{-1}^{1} t\, l_i^n(t) l_j^n(t) \frac{dt}{2} = t_i^n \beta_i^n \delta_{i,j}, \quad \text{où} \quad \beta_i^n = \frac{1}{2} \int_{-1}^{1} (l_i^n)^2 dt. \tag{3.21}$$

Puisque a depend linéairement de y comme dans (1.7), alors en formulant le système de Galerkin avec les polynômes $\{l_\nu^\mu\}_{\nu \in \mathcal{B}_\mu}$, il est facile de voir que les coordonnées correspondantes $u_{\mathcal{B}_\mu,\nu}$ de la projection de Galerkin $u_{\mathcal{B}_\mu}$, sont chacunes l'unique solution dans $H_0^1(D)$ du problème variationnel suivant

$$\beta_\nu^\mu \int_D (\overline{a} + \sum_{j=1}^{d} t_{\nu_j}^{\mu_j+1} \psi_j) \nabla u_{\mathcal{B}_\mu,\nu} \nabla w = \int_U \int_D f w l_\nu^\mu d\varrho, \qquad w \in H_0^1(D), \tag{3.22}$$

où $\beta_\nu^\mu = \prod_{j=1}^d \beta_{\nu_j}^{\mu_j+1}$. Le calcul de la projection de Galerkin revient donc à résoudre $\prod_{j=1}^d (1+\mu_j)$ problèmes aux limites déterministes, équivalents en coût au calcul d'une instance de l'application solution u. En outre, soulignons la remarque suivante qui n'est pas mentionnée dans [5]. Si f ne dépend pas de y, alors le terme de droite dans (3.22) est le produit de deux intégrales et puisque les solutions $u(y)$ de (1.6) satisfont les problèmes suivants

$$\int_D (\overline{a} + \sum_{j=1}^{d} y_j \psi_j) \nabla u(y) \nabla w = \int_D f w, \qquad w \in H_0^1(D), \tag{3.23}$$

alors en notant $t_\nu^\mu := (t_{\nu_1}^{\mu_1+1}, \ldots, t_{\nu_d}^{\mu_d+1}) \in U$, on obtient que

$$u_{\mathcal{B}_\mu,\nu} = \frac{w_\nu^\mu}{\beta_\nu^\mu} u(t_\nu^\mu), \qquad w_\nu^\mu := \int_U l_\nu^\mu(y) d\varrho(y) = \prod_{j=1}^{d} \int_U l_{\nu_j}^{\mu_j+1}(t) \frac{dt}{2} \tag{3.24}$$

Notons que w_ν^μ est le poids de Gauss associé avec l'abscisse multi-dimensionnelle t_ν^μ pour la quadrature dans la grilles de points $\{t_\nu^\mu : \nu \in \mathcal{B}_\mu\}$.

Dans un travail ultérieur [4], les auteurs proposent de calculer une approximation de u dans l'espace $\mathbb{V}_{\mathcal{B}_\mu}$ directement par la collocation du problème variationel (3.23) satisfait par les fonctions $u(y) \in V$, sur la grille tensorisée

$$\Gamma_{\mathcal{B}_\mu} := \{t_\nu^\mu : \nu \leq \mu\} = \otimes_{j=1}^d \{t_0^{\mu_j+1}, \ldots, t_{\mu_j}^{\mu_j+1}\}, \tag{3.25}$$

puis en construisant l'approximation par interpolation. Grâce à la structure produit de $\mathbb{P}_{\mathcal{B}_\mu}$, il est facile de voir que les $\{l_\nu^\mu\}_{\nu \in \mathcal{B}_\mu}$ sont les polynômes de Lagrange associés avec la grille $\Gamma_{\mathcal{B}_\mu}$ et l'espace $\mathbb{P}_{\mathcal{B}_\mu}$. Par conséquent, l'opérateur d'interpolation est donné par

$$\mathcal{I}_\mu u := \sum_{\nu \leq \mu} u(t_\nu^\mu) l_\nu^\mu. \tag{3.26}$$

Une façon d'analyser la stabilité de l'opérateur d'interpolation consiste à exploiter la propriété de double orthogonalité (3.21). Plus précisément, avec $\|u\|_{\mathcal{V}_\infty} := \sup_{y\in U} \|u(y)\|_V$, on a

$$\mathbb{E}[\|\mathcal{I}_\mu u(y)\|_V^2] = \sum_{\nu\le\mu} \|u(t_\nu^\mu)\|_V^2 \int_U (l_\nu^\mu(y))^2 d\varrho(y) \le \|u\|_{\mathcal{V}_\infty}^2 \sum_{\nu\le\mu} \int_U (l_\nu^\mu(y))^2 d\varrho(y) = \|u\|_{\mathcal{V}_\infty}^2, \tag{3.27}$$

où la dernière identité découle de l'orthogonalité (3.21), et le fait que $\sum_{\nu\le\mu} l_\nu^\mu(y) = 1$. Par conséquent,

$$\mathbb{E}[\|u - \mathcal{I}_\mu u\|_V^2] \le 2 \inf_{v\in\mathbb{V}_{\mathcal{B}_\mu}} \sup_{y\in U} \|u(y) - v(y)\|_V^2. \tag{3.28}$$

Par un examen approfondi de la croissance des dérivées partielles de u comme dans [5], les auteurs montrent que l'application u admet une extension holomorphe dans le domaine complexe et utilisent cette propriété pour montrer que le terme de droite dans la dernière inégalité satisfait une borne similaire à l'erreur L^2 dans (3.18).

Dans [70, 69], les approximations polynomiales sont construites avec les méthodes de collocations dans des espaces de polynômes qui ne sont pas nécessairement de type produit tensoriel, suivant l'approche des *sparse grids* initialement due à Smolyak [76] et étudiée dans plusieurs travaux, entre autres [50, 71, 9, 81]. Dans [70], ces approximations polynomiales sont considérées dans des espaces *isotropes* $\mathbb{V}_{m(\mathcal{S}_k)}$, où m est une function de $\mathbb{N}$ dans $\mathbb{N}$ donnée croissante, qui satisfait $m(0) = 0$ et la convention que $m(-1) = -1$, $\mathcal{S}_k$ est le simplexe dans (3.13) et la notation $m(\mathcal{S}_k)$ signifie

$$m(\mathcal{S}_k) := \bigcup_{i\in\mathcal{S}_k} B_{m(i)} \quad \text{avec} \quad B_{m(i)} := \Big\{\nu \in \mathbb{N}^d : m(i_j - 1) < \nu_j \le m(i_j)\Big\}. \tag{3.29}$$

Notons que $m(\mathcal{S}_k)$ est un ensemble fermé vers le bas et qu'il coincide avec $\mathcal{S}_k$ lorsque m est la fonction identité. L'approximation polynomiale est donnée par la formule de Smolyak

$$\mathcal{I}_{m(\mathcal{S}_k)} u = \sum_{i\in\mathcal{S}_k} \otimes_{j=1}^d (I_{m(i_j)} - I_{m(i_j-1)})u, \tag{3.30}$$

où $I_{-1} := 0$, et pour tout $l \ge 0$, $I_{m(l)}$ l'opérateur d'interpolation de Lagrange associé avec $m(l) + 1$ points distincts $\{r_0, \dots, r_{m(l)}\}$ dans $[-1, 1]$. Lorsque les points d'interpolation des opérateurs $I_{m(0)}, I_{m(1)}, \dots$ sont les sections emboitées d'une suite infinie $r_0, r_1, r_2 \dots$ de points mutuellement distincts, l'opérateur $\mathcal{I}_{m(\mathcal{S}_k)}$ est un opérateur d'interpolation associé avec l'espace $\mathbb{V}_{m(\mathcal{S}_k)}$ et la grille parcimonieuse isotrope de points de U

$$\Gamma_{m(\mathcal{S}_k)} := \Big\{r_\nu := (r_{\nu_1}, \dots, r_{\nu_d}) : \nu \in m(\mathcal{S}_k)\Big\}, \tag{3.31}$$

voir [7, 9, 26]. Le schéma que nous avons décrit soulève le problème de l'optimisation du compromis entre la croissance de m qui dicte le nombre total de points de collocation et le choix des positions de ces points qui détermine la qualité de l'approximation. Dans le

cas des sections emboitées, la qualité de l'approximation est étudiée via la stabilité de l'opérateur d'interpolation, quantifiée par sa constante de Lebesgue. Dans [70], le choix classique des points de Clenshaw-Curtis emboités est examiné, plus précisément le choix de $m(l)$ et des sections emboitées des points d'interpolation associés aux opérateurs $I_{m(l)}$ est donné par

$$m(l) = 2^l, \quad l \geq 1, \quad \text{et} \quad \{r_0, \ldots, r_{m(l)}\} = \Big\{ \cos\Big(\frac{j}{2^l}\pi\Big) : j = 0, \ldots, m(l) \Big\}. \tag{3.32}$$

Il est connu que les opérateurs $I_{m(l)}$ ont des constantes de Lebesgue qui croissent comme $\log(2^l)$. Des résultats de convergence sont obtenus en combinant cette dernière propriété et les résultats d'analyticité obtenus dans [5].

Les espaces de polynômes $\mathbb{V}_{m(\mathcal{S}_k)}$ ci-dessus sont isotropes, et par conséquent les approximations $\mathcal{I}_{m(\mathcal{S}_k)}u$ sont aussi isotropes, dans le sens où les variables y_j jouent des rôles symétriques. Pour certains problèmes, la solution u a une dépendance fortement anisotrope en les variables individuelles y_j, par exemple quand les fonctions ψ_j dans (1.7) ont des normes $\|\psi_j\|_{L^\infty(D)}$ qui varient fortement avec j. Il convient alors de choisir des approximations polynomiales qui reflètent cette anisotropie. Dans [69], les auteurs traitent ce problème en considérant des versions anisotropes de l'espace $\mathbb{V}_{\mathcal{S}_k}$ et par conséquent des versions anisotropes des $\mathbb{V}_{m(\mathcal{S}_k)}$. Ces versions sont caractérisées par des paramètres $\alpha := (\alpha_1, \ldots, \alpha_d) \in {\mathbb{R}_+^*}^d$ suivant

$$\mathcal{S}_{k,\alpha} := \Big\{ \nu \in \mathbb{N}^d : \langle \nu, \alpha \rangle := \sum_{j=1}^{d} \nu_j \alpha_j \leq k \Big\}, \qquad m(\mathcal{S}_{k,\alpha}) := \bigcup_{i \in \mathcal{S}_{k,\alpha}} B_{m(i)}. \tag{3.33}$$

Notons à nouveau que ces ensembles sont fermés vers le bas. Le paramètre α doit refléter l'anisotropie du problème: plus la dépendance en la variable y_j est faible, plus la valeur de α_j est grande. Dans la cas isotrope, toutes les coordonnées de α sont égales à 1. Les approximations considérées dans [69] sont construites avec la même formule de Smolyak (3.30), avec maintenant $i \in \mathcal{S}_{k,\alpha}$. Comme dans le cas isotrope, lorsqu'il y a emboitement des points, l'opérateur est un opérateur d'interpolation associé à l'espace $\mathbb{V}_{m(\mathcal{S}_{k,\alpha})}$ et la grille parcimonieuse anisotrope de points

$$\Gamma_{m(\mathcal{S}_{k,\alpha})} := \Big\{ r_\nu := (r_{\nu_1}, \ldots, r_{\nu_d}) : \nu \in m(\mathcal{S}_{k,\alpha}) \Big\}, \tag{3.34}$$

voir [7, 26]. Comme dans [70], l'analyse de la convergence est basée sur la stabilité de l'opérateur d'interpolation et les résultats d'analyticité de [5]. Il est en particulier prouvé que lorsque l'on utilise les points de Clenshaw-Curtis, il existe un choix optimal de α dépendant des rayons d'analyticité de u dans chaque variable y_j, pour lequel les bornes d'erreur sont minimales. Les méthodes de collocation de type "sparse grid" sont ensuite étendues à des espaces de polynômes encore plus anisotropes dans [7, 8], où cette fois les ensembles Λ sont construits de façon adaptative et optimisés à l'aide de l'algorithme knapsack.

Toutes les stratégies décrites ci-dessus produisent des approximations polynomiales calculables qui convergent vers l'application solution u dans les deux sens uniforme et/ou quadratique moyen. Afin de comparer leurs performance numériques, un benchmark adéquat est l'analyse de la décroissance de l'erreur en fonction du coût numérique total. Pour les méthodes que nous avons présentées jusqu'à maintenant, le coût du calcul de l'approximation dans un espace donnné $\mathbb{V}_\Lambda$ est essentiellement dominé par le coût de $\#(\Lambda)$ évaluations de u. Il est donc pertinent d'étudier l'erreur de chaque méthode comme une fonction de $n := \#(\Lambda)$. Ceci n'est pas immédiat, car les erreurs sont pour la plupart des stratégies données sous forme de sommes de contributions d'erreur. Cependant, un examen détaillé révèle que pour chaque méthode, les vitesses de convergences ne sont pas meilleures que

$$\xi^{\lambda(\Lambda)}, \qquad \lambda(\Lambda) := \max_{\nu\in\Lambda}\{\nu_{\max} : \nu_{\max} := \max(\nu_1,\dots,\nu_d)\}, \tag{3.35}$$

où ξ un nombre donné dans $]0,1[$ indépendent de d. Notons que $\lambda(\Lambda)$ est le degré maximal atteint en au moins une variable pour les polynômes de $\mathbb{P}_\Lambda$. Pour les approximations isotropes, les cardinalités des ensembles d'indices considérés sont $\#(\mathcal{B}_\mu) = (1+\mu_1)^d$ avec $\mu := (\mu_1,\dots,\mu_1)$, $\#(\mathcal{S}_k) = \binom{k+d}{k}$ et avec m la function de doublement de (3.32), on a si $d \geq k$,

$$\#(m(\mathcal{S}_k)) = \sum_{i\in\mathcal{S}_k}\#(B_{m(i)}) \geq \sum_{i\in\mathcal{S}_k\cap\{0,1\}^d} 2^{|i|} = \sum_{j=0}^{k}\binom{d}{j}2^j \geq d2^k. \tag{3.36}$$

Comme $\lambda(\mathcal{B}_\mu) = 1+\mu_1$, $\lambda(\mathcal{S}_k) = k$ et $\lambda(m(\mathcal{S}_k)) = 2^k$, alors il est facilement vérifiable que

$$\lambda(\mathcal{B}_\mu) \leq (\#(\mathcal{B}_\mu))^{\frac{1}{d}}, \qquad \lambda(\mathcal{S}_k) \leq \frac{k!}{d^k}\#(\mathcal{S}_k), \qquad \lambda(m(\mathcal{S}_k)) \leq \frac{1}{d}\#(m(\mathcal{S}_k)), \tag{3.37}$$

et donc au vu de (3.35), les taux des approximations isotropes se détériorent avec la croissance de la dimension d et ne sont donc pas robustes aux grandes dimensions. L'approximation dans des espaces anisotropes permet dans une certaine mesure de contourner cette restriction car la dimension d peut être réduite en activant seulement peu de variables y_j dans l'approximation. Par example, poser $\mu_j = 0$ lorsque l'ensemble $\mathcal{B}_\mu$ est considéré ou prendre $\alpha_j >> 1$ lorsque les ensembles $\mathcal{S}_{k,\alpha}$ et $m(\mathcal{S}_{k,\alpha})$ sont considérés, conduit à l'inactivité de la variable y_j dans l'approximation. Cependant, à cause de leurs formes rigides, les ensembles anisotropes doublent au minimum en cardinalité lorsqu'une nouvelle coordonnée y_j est activée alors que seulement la j-ème contribution de l'erreur décroît. Signalons aussi que pour les grilles parcimonieuses (sparse grids), les constantes multiplicatives dans les bornes sur les erreurs dépendent de d.

À la lumière de la discussion précédente, un objectif consiste donc à concevoir des méthodes polynomiales qui ont des taux de convergence robustes à la dimension paramétrique d. Un objectif équivalent est de concevoir des méthodes polynomiales

avec des taux de convergence dans le cas de la dimension infinie $d = \infty$, par exemple des taux algébriques, c'est à dire des bornes d'erreur de la forme Cn^{-s}, où $n = \#(\Lambda)$. Bien entendu, un tel objectif ne peut être atteint si toutes les variables y_j pèsent de façon identique dans les variations de la solution u pour les raisons expliquées précédemment. Par ailleurs, la convergence d'un développement de type (1.11) exige une certaine décroissance des ψ_j ce qui justifie que les variables y_j sont moins actives quand $j \to +\infty$.

Dans [34, 33], un nouveau paradigme a été introduit pour traiter le modèle (1.6) dans le cas $d = \infty$ et atteindre les objectifs mentionnés ci-dessus. Le résultat suivant est démontré: si a est de la forme (1.11) avec la suite $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$ appartenant à $\ell^p(\mathbb{N})$ pour une valeur de $p < 1$, alors les séries de Taylor

$$\sum_{\nu\in\mathcal{F}} t_\nu y^\nu \quad \text{avec} \quad t_\nu := \frac{\partial_\nu u(0)}{\nu!} \in V, \tag{3.38}$$

restreintes à des ensembles bien choisis $(\Lambda_n)_{n\geq 1}$, avec $\#(\Lambda_n) = n$, convergent vers u au sens uniforme avec la vitesse algébrique n^{-s}, à une constante multiplicative près ne dépendant que de $\|(\psi_j)\|_{\ell^p(\mathbb{N})}$, avec

$$s := \frac{1}{p} - 1. \tag{3.39}$$

Le même résultat est démontré pour les séries de Legendre

$$\sum_{\nu\in\mathcal{F}} v_\nu L_\nu \quad \text{avec} \quad v_\nu := \int_U u(y) L_\nu(y) d\varrho(y) \in V, \tag{3.40}$$

au sens quadratique moyen, avec la meilleure vitesse n^{-s^*} où

$$s^* := \frac{1}{p} - \frac{1}{2}. \tag{3.41}$$

En d'autres termes, on peut établir des vitesses de convergence algébriques qui sont robustes à la croissance de la dimension d, sous une hypothèse de décroissance des fonctions ψ_j quand $j \to +\infty$ reflétant l'anisotropie de l'application solution.

Afin d'établir les résultats précédents, l'idée de base consiste à définir les ensembles Λ_n comme étant ceux des indices associés avec les n plus grands termes des suites $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ et $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ des coefficients de Taylor et Legendre. Cette troncation est une forme d'approximation non-linéaire de l'application solution, souvent appelée *meilleure approximation à n termes*, voir [40]. Les ensembles Λ_n qui en résultent peuvent être assez différents des ensembles isotropes et anisotropes que nous avons décrit précédemment.

La connaissance exacte de la meilleure approximation à n termes est souvent inaccessible, et par conséquent les résultats précédents doivent être vus comme des résultats

d'approximation théoriques. En particulier, ils ne débouchent pas naturellement sur des stratégies numériques. Soulignons aussi que les preuves des résultats ci-dessus exploitent fortement la nature linéaire elliptique du modèle (1.6) et la dépendance paramétrique affine dans (1.11). Ces limitations soulèvent les questions suivantes:

(i) Peut-on construire des ensembles d'indices $(\Lambda_n)_{n\geq 1}$ avec des stratégies numériques adaptatives où non-adaptatives, tels que le taux de convergence des approximations polynomiales de l'application solution dans l'espace $\mathbb{V}_{\Lambda_n}$ est similaire à celui de la meilleure approximation à $\#(\Lambda_n)$ termes ?

(ii) Étant donné ces ensembles d'indices, peut on définir les polynômes d'approximation de l'application solution par des stratégies numériques simples, par exemple par des méthodes non-intrusives telle que l'interpolation, et obtenir un taux de convergence similaire?

(iii) Peut-on obtenir des résultats d'approximation similaires pour des modèles plus généraux, y compris des EDP non-linéaires avec une dépendance non-affine des paramètres, et toujours en la dimension paramétrique $d = \infty$?

Cette thèse est motivée par ces questions et elle apporte des réponses précises à chacune d'entre elles.

4 Plan de la thèse

Cette thèse est constituée de sept chapitres et contient trois parties principales numérotées I, II et III. La plupart des résultats présentés dans cette thèse sont publiés dans nos articles [22, 26, 25, 21, 24, 23].

La partie I traite des résultats théoriques d'approximation, alors que les parties II et III traitent de la construction d'algorithmes pratiques d'approximation. Dans la partie I (chapitres 1-2), nous rappelons les résultats de [34] et [33] pour les EDP elliptiques linéaires (1.6) avec dépendance paramétrique affine (1.7) et nous présentons nos résultats de [25] qui s'appliquent à des modèles plus généraux tels que des EDP non-linéaires avec une dépendance paramétrique non-affine. Dans la partie II (chapitres 3-4) nous présentons deux algorithmes intrusifs pour l'approximation de l'application solution u de (1.6), l'un basé sur des séries de Taylor construites de façon adaptative suivant l'approche adoptée dans [22], l'autre basé sur des projections de Galerkin dans la lignée des méthodes de [53]. Dans la partie III (chapitres 5-6-7), nous présentons deux algorithmes non-intrusifs qui peuvent être utilisés pour des EDP paramétriques plus générales : l'interpolation et les moindres carrés, suivant les approches que nous avons développés dans [26, 21, 24] et dans [23]. Il convient de noter que tout au long

de cette thèse, nous nous plaçons dans le cadre de la dimension infinie,

$$d = \infty \qquad \text{et} \qquad U := [-1,1]^{\mathbb{N}}, \tag{4.1}$$

mais que les différents algorithmes et les résultats sont aussi applicables dans le cadre de la dimension finie $d < \infty$.

Dans le chapitre 1, nous étudions le problème paramétrique elliptique (1.6) avec la dépendance paramétrique affine (1.7) et l'hypothèse d'ellipticité uniforme (1.8). Pour ce modèle, nous rappelons en détail les résultats d'approximation obtenus dans [34, 33]. Comme nous l'avons expliqué précédemment, ces résultats montrent que sous l'hypothèse $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathcal{F})$ pour un certain $p < 1$, les series de Taylor restreintes aux ensembles Λ_n de leurs n plus grands termes convergent vers l'application solution u avec une vitesses algébrique n^{-s} où $s := \frac{1}{p} - 1$ dans le sens uniforme, et celles de Legendre convergent avec la vitesse n^{-s^*} où $s^* := \frac{1}{p} - \frac{1}{2}$ dans le sens quadratique moyen. En outre, nous montrons que les ensembles $(\Lambda_n)_{n\geq 0}$ des meilleures approximations à n termes utilisés dans ces résultats peuvent être modifiés pour être fermés vers le bas dans le sens de (3.8), sans détérioration de la vitesse de convergence. Ceci est d'une grande importance dans l'étude de la construction et la convergence des algorithmes numériques des parties II et III. Bien que le chapitre 1 réunisse dans une certaine mesure les résultats obtenus dans [34, 33], il est auto-contenu et peut être lu sans avoir recours aux papiers cités. Nous avons souhaité expliquer en détail le paradigme du traitement de la dimension infinie, tout en raccourcissant et simplifiant les raisonnements de [34, 33]. L'analyse du chapitre 1 est essentielle pour la compréhension des divers outils que nous utilisons dans les chapitres suivants.

Dans le chapitre 2 qui reprend notre article [25], nous étudions des EDP paramétriques de la forme générale (1.1) avec des dépendances anisotropes en les paramètres y_j. Ces EDP ne sont pas nécessairement de type (1.6) avec dépendance affine comme dans (1.7). Pour donner un exemple simple, considérons le modèle (1.6) avec le coefficient de diffusion

$$a(x,y) := \overline{a} + \Big(\sum_{j\geq 1} y_j \psi_j\Big)^2. \tag{4.2}$$

Bien que l'ellipticité uniforme en $y \in U$ est maintenue pour cette nouvelle forme de a, les séries de Taylor peuvent ne pas converger. Ceci est déjà le cas quand toutes les fonctions ψ_j sont nulles sauf $\psi_1 = b > 1$ constante et $\overline{a}$ constant égal à 1. En effet, dans ce cas

$$u(y) = \frac{u(0)}{1 + b^2 y_1^2}, \qquad y \in U, \tag{4.3}$$

est une fonction dont la série de Taylor diverge sur $[-1,1]$. En revanche, cette fonction reste la somme de sa série de Legendre. En suivant l'approche du chapitre 1 pour l'approximation par les polynômes de Legendre, nous montrons qu'une large classe d'EDP paramétriques peut être approchée avec des vitesses de convergence algébriques

n^{-s} par les meilleures approximations à n-termes associées aux séries de Legendre. Pour faire cela, nous introduisons la notion de (p,ε)-holomorphie qui décrit certaines hypothèses d'anisotropie pouvant garantir la vitesse de convergence algébrique. On introduit deux cadres généraux dans lesquels la (p,ε)-holomorphie est satisfaite, le premier basé sur des conditions Inf-Sup et le second sur la version holomorphe du théorème des functions implicites dans des espaces de Banach. Nous montrons que ces deux cadres s'appliquent à diverses EDP paramétriques. En particulier, nous considérons des problèmes elliptiques semi-linéaires de la forme

$$g(u) - \operatorname{div}(a(y)\nabla u) = f, \tag{4.4}$$

des problèmes paraboliques tels que

$$\partial_t u - \operatorname{div}(a(y)\nabla u) = f, \tag{4.5}$$

et des EDP considérées sur des domaines qui dépendent du paramètre y.

Le reste de la thèse, parties II et III, est consacré à la conception d'algorithmes qui permettent en pratique de construire des approximations polynomiales pour les EDP paramétriques. Étant donnée une famille de polynômes $\mathcal{P}$ du type (3.3), deux problèmes principaux doivent être traités:

- L'identification de bons ensembles d'indices Λ_n dans le sens où u peut être bien approchée dans $\mathbb{V}_{\Lambda_n}(\mathcal{P})$.
- Le calcul d'une bonne approximation $u_{\Lambda_n} \in \mathbb{V}_{\Lambda_n}(\mathcal{P})$ de u.

Comme nous l'avons souligné à plusieurs reprises, il est d'une grande importance que les deux tâches soient raisonnables d'un point de vue du coût de calcul. Nous verrons que pour l'approximation, tant au sens uniforme qu'au sens quadratique moyen, il n'est pas sans intérêt de considérer des algorithmes adaptatifs. Pour ce type d'algorithmes, la suite des ensembles d'indices $(\Lambda_n)_{n\geq 1}$ n'est pas fixée d'avance, l'identification de l'ensemble Λ_{n+1} étant basée sur l'information dont on dispose à l'issue du calcul à l'étape n. Nous montrons de façon rigoureuse l'efficacité de ce type d'algorithmes dans les chapitres 3-4 qui traitent du problème elliptique (1.6).

Le chapitre 3 reprend notre article [22] consacré à l'approximation de u par des séries de Taylor numériquement calculables. Dans [34, 33], il est prouvé que les coefficients de Taylor t_ν, définis dans (3.38), sont les uniques solutions des problèmes récursifs suivants

$$\int_D \overline{a}(x)\nabla t_\nu(x)\nabla w(x)dx = -\sum_{j:\nu_j\neq 0}\int_D \psi_j(x)\nabla t_{\nu-e_j}(x)\nabla w(x)dx, \qquad w\in V, \tag{4.6}$$

où $e_j = (\delta_{i,j})_{i\geq 1}$ est la suite de Kronecker d'indices j et $\nu - e_j$ la soustraction vectorielle de e_j à ν, i.e.

$$\nu - e_j = (\nu_1, \ldots, \nu_{j-1}, \nu_j - 1, \nu_{j+1}, \ldots). \tag{4.7}$$

En se basant sur cette récurrence et en utilisant des arguments de réduction de résidu dans la lignée de ceux introduits pour les méthodes adaptatives d'ondelettes [30, 31], nous montrons qu'il est possible de construire de proche en proche une suite d'ensembles emboités $(\Lambda_k)_{k\geq 1}$ tels que les séries de Taylor associées convergent avec la vitesse optimale n^{-s}, où $n = n(k) := \#(\Lambda_k)$, établie dans le chapitre 1. Au vu de (4.6), la série de Taylor associée à un ensemble d'indices Λ donné peut être calculée en exactement $\#(\Lambda)$ étapes de récurrence si et seulement si

$$\nu \in \Lambda \Rightarrow \nu - e_j \in \Lambda \text{ pour tout } j \geq 1 \text{ tel que } \nu_j \neq 0. \tag{4.8}$$

Cette définition est équivalente à (3.8). Les ensembles d'indices qui sont construits de façon adaptative pour les séries de Taylor sont donc toujours fermés vers le bas.

Dans le chapitre 4, nous étudions l'approximation de u par des projections de Galerkin dans le sens quadratique moyen, à l'aide des polynômes de Legendre $(L_\nu)_{\nu\in\mathcal{F}}$. Nous considérons seulement le cas où la probabilité jointe de y est la mesure uniforme sur U. Cependant, tous les résultats s'étendent de manière immédiate à d'autres mesures produit, en remplaçant les polynômes de Legendre par les polynômes tensorisés orthogonaux pour la mesure produit. En suivant l'approche de [53], nous formulons le problème variationnel (3.14) dans la base de Legendre, ce qui conduit à une formulation matricielle

$$\mathbf{A}\mathbf{u} = \mathbf{f} \tag{4.9}$$

où $\mathbf{A} = (\mathbf{A}_{\nu,\nu'})_{\nu,\nu'\in\mathcal{F}}$ est une matrice infinie d'opérateurs définis de V dans V^*, $\mathbf{u} = (v_\nu)_{\nu\in\Lambda} \in \ell^2(\mathcal{F}, V)$ la suite des coefficients de Legendre de u et $\mathbf{f} = (f_\nu)_{\nu\in\Lambda} \in \ell^2(\mathcal{F}, V^*)$ la suite des coefficients de Legendre de $f \in (L^2(U, V, d\varrho))^*$. En suivant une approche analogue à celle des méthodes adaptatives en ondelettes pour les opérateurs elliptiques [30, 31, 49], nous étudions les propriétés de la matrice infinie $\mathbf{A}$, puis par une analyse de résidus, nous construisons de proche en proche une suite d'ensembles emboités $(\Lambda_k)_{k\geq 1}$ tels que la projection de Galerkin $\mathbf{u}_{\Lambda_k} \in \ell^2(\Lambda_k, V)$ de la formulation précédente converge vers $\mathbf{u}$ avec la vitesse n^{-s^*}, où $n = n(k) := \#(\Lambda_k)$, établie dans le chapitre 1. Contrairement au chapitre 3, les ensembles Λ_k ne sont pas nécessairement fermés vers le bas. Nous montrons que des résultats d'approximation similaires peuvent être obtenus avec les ensembles fermés vers le bas, et que dans ce cas, les projections de Galerkin approchent aussi u dans le sens uniforme.

Les méthodes que nous présentons dans les chapitres 3-4 sont intrusives. Elles sont spécifiquement conçues pour le problème elliptique linéaire (1.6) avec dépendance affine comme dans (1.7). En particulier, l'analyse de la convergence est fortement liée à ces spécificités. Pour des problèmes plus généraux, ces méthodes sont difficiles à mettre en oeuvre, et on peut donc leur préférer des méthodes non-intrusives. Celles-ci deviennent incontournables dans les cas où on n'a pas une connaissance complète de l'EDP et on peut seulement obtenir les solutions $u(y)$ pour tout paramètre y via un solveur numérique. Dans cette perspective, nous étudions dans la partie III deux méthodes non-intrusives souvent utilisées : l'interpolation et les moindres carrés.

Dans le chapitre 5, nous présentons le schéma d'interpolation que nous avons introduit dans [26]. Le schéma est défini par la généralisation de la formule de Smolyak qui donne les opérateurs d'interpolation (3.30) pour les simplexes isotropes et anisotropes $\mathcal{S}_k$ et $\mathcal{S}_{k,\alpha}$, maintenant remplacés par des ensembles fermés vers le bas quelconques. Nous généralisons ainsi les résultats des articles [9, 7] dans lesquels il est prouvé que pour certains types d'ensembles fermés vers le bas Λ, l'opérateur $\mathcal{I}_\Lambda$ defiué par (3.30) avec simplement Λ au lieu de $\mathcal{S}_k$ est un opérateur d'interpolation. Plus précisement, étant donné $(r_0, r_1, r_2, \ldots)$ une suite de points deux à deux distincts dans $[-1, 1]$ et en désignant par I_k l'opérateur d'interpolation polynomiale associé avec $(r_0, \ldots, r_k)$, avec la convention $I_{-1} = 0$, alors pour tout ensemble fermé vers le bas Λ,

$$\mathcal{I}_\Lambda := \sum_{i\in\Lambda} \otimes_{j=1}^{d} (I_{i_j} - I_{i_j-1}), \tag{4.10}$$

est un opérateur d'interpolation sur $\mathbb{P}_\Lambda$ pour la grille de points

$$\Gamma_\Lambda := \Big\{ r_\nu := (r_{\nu_j})_{j\geq 1} : \nu \in \Lambda \Big\}. \tag{4.11}$$

Nous montrons que ces opérateurs peuvent être calculés de façon simple par une formule de type Newton. Plus précisément, étant donné Λ un ensemble fermé vers le bas et $\nu \in \mathcal{F} \setminus \Lambda$ telle que $\Lambda' := \Lambda \cup \{\nu\}$ est fermé vers le bas, alors

$$\mathcal{I}_{\Lambda'} u = \mathcal{I}_\Lambda u + \Big(u(r_\nu) - \mathcal{I}_\Lambda u(r_\nu) \Big) h_\nu, \tag{4.12}$$

où

$$h_\nu(y) = \prod_{j\geq 1} h_{\nu_j}(y_j), \quad \text{avec} \quad h_0 = 1 \quad \text{et} \quad h_k(t) := \prod_{j=0}^{k-1} \frac{t - r_j}{r_k - r_j}. \tag{4.13}$$

Nous étudions ensuite la stabilité de l'interpolation via l'analyse de la constante de Lebesgue $\mathbb{L}_\Lambda := \|\mathcal{I}_\Lambda\|_{L^\infty \to L^\infty}$. Nous montrons en particulier que la croissance de ces constantes en fonction de la taille Λ peut être estimée à partir de la croissance des constantes de Lebesgue $\mathbb{L}_k$ associées aux opérateurs I_k. Plus précisement, nous montrons que

$$\mathbb{L}_k \leq (k+1)^\theta, \quad k \geq 1 \quad \Rightarrow \quad \mathbb{L}_\Lambda \leq (\#\Lambda)^{\theta+1} \quad \text{pour tout ensemble bas } \Lambda. \tag{4.14}$$

La croissance polynomiale $(\#(\Lambda))^{\theta+1}$ peut être plus grande que la décroissance algébrique $(\#(\Lambda))^{-s}$ que nous avons établie dans les chapters 1-2 pour l'approximation des EDP dans les espaces de polynômes $\mathbb{V}_\Lambda$. Cependant, nous montrons que sous les mêmes hypothèses que celles des chapitres 1-2, il existe une suite d'ensembles fermés vers le bas $(\Lambda_n)_{n\geq 0}$ avec $\#(\Lambda_n) = n$, tel que l'approximation de u par $\mathcal{I}_{\Lambda_n} u$ converge avec la vitesse optimale n^{-s} où $s = \frac{1}{p} - 1$.

Nous utilisons également la formule (4.12) comme point de départ pour le développement d'algorithmes adaptatifs où, pour Λ_n donné, l'indice selectionné ν tel que $\Lambda_{n+1} =$

$\Lambda_n \cup \{\nu\}$ maximise pour une norme d'intérêt fixé l'incrément $\Big(u(r_\nu) - \mathcal{I}_\Lambda u(r_\nu)\Big)h_\nu$. Bien que la convergence de tels algorithmes adaptatifs avec une vitesse optimale n'est pas assurée, ils présentent des comportements numériques satisfaisants dans plusieurs cas tests. Finalement, nous étendons l'idée de l'interpolation parcimonieuse en grande dimension dans le cas de systèmes tensorisés autres que les polynômes. En particulier, pour les fonctions affines par morceaux et quadratiques par morceaux, en introduisant des concepts similaires d'ensemble fermés vers le bas.

Le résultat (4.14) motive la recherche d'une suite infinie $(r_0, r_1, r_2, \ldots)$ telle que les constantes de Lebesgue $\mathbb{L}_k$ associées aux sections $(r_0, \ldots, r_k)$ ont une croissance modérée. Notons que les constantes de Lebesgue des abscisses de Chebychev ou de Gauss-Lobatto ont des croissances logarithmiques, cependant ces points ne sont pas les sections d'une suite infinie de points. Dans le chapitre 6, nous étudions la croissance de la constante de Lebesgue associée avec les suites dites de Leja sur le disque unité complexe et leurs projections sur $[-1, 1]$ appelées suites de $\Re$-Leja. Ce chapitre améliore les résultats de notre article [21] et de deux autres travaux antérieurs [18, 19] dans lesquels ce type de suite est étudié. Nous donnons de nouvelles propriétés structurelles de ces suites, puis nous établissons une nouvelle borne sur la croissance de la constante de Lebesgue des suites de $\Re$-Leja. Plus précisément, nous montrons que

$$\mathbb{L}_k \leq 8\sqrt{2}(k+1)^2, \quad k \geq 0, \tag{4.15}$$

ce qui améliore la borne $8(k+1)^2 \log(k+1)$ que nous avons établi dans [21]. Ce nouveau résultat montre en particulier que, étant donné une suite de $\Re$-Leja $(r_0, r_1, r_2, \ldots)$, les opérateurs d'interpolation en grande dimension $\mathcal{I}_\Lambda$ qui lui sont associés ont des constantes de Lebesgue bornées par $(\#(\Lambda))^3$ quelque soit la dimension d et la forme de l'ensemble Λ.

Dans le chapitre 7, nous présentons les résultats de notre article [23] dans lequel nous étudions la stabilité de la méthode des moindres carrés par les polynômes en grande dimension. Étant donné Λ un ensemble fermé vers le bas de cardinalité n et $\mathcal{O}_m := (y^i, z^i)_{i=1,\ldots,m}$, où les y^i sont des réalisations indépendantes du paramètre vectoriel aléatoire y et z^i sont des observation bruités ou non-bruités de l'application solution en y^i, la projection des moindres carrés est définie par

$$\mathcal{I}_{\Lambda,\mathcal{O}_m} u := \operatorname*{argmin}_{v \in \mathbb{V}_\Lambda} \frac{1}{m} \sum_{i=1}^{m} \|z^i - v(y^i)\|_V^2. \tag{4.16}$$

Lorsque V est un espace de Hilbert, la solution du problème est obtenue par la résolution d'un système linéaire simple similaire au cas de données à valeurs réelles ou complexes. En se basant sur les techniques introduites dans [32], nous examinons la stabilité des projections des moindres carrés, en faisant apparaître un compromis entre la dimension $n = \#(\Lambda)$ de l'espace de polynômes $\mathbb{V}_\Lambda$ et la taille m de l'échantillon. En particulier, lorsque ϱ est la mesure uniforme sur U, nous montrons que la projection est stable lorsque m est au moins de l'ordre de n^2.

Introduction: English version

1 Parametric partial differential equations

This thesis is devoted to the theoretical study and numerical approximation of high dimensional *parametric* partial differential equations (PDEs). Parametric PDE's appear in various context for modeling dependence of a specified physical phenomenon with respect to cerain relevant parameters. For example, the heat distribution on a steel plate for various type of steel, the parametrization being on the percentage of the different chemical elements constituting the alloy. The abstract representation that we shall adopt for parametric PDEs is

$$\mathcal{D}(u, y) = 0, \tag{1.1}$$

where $\mathcal{D}$ is a linear or nonlinear partial differential operator, modelling the physical phenomenon, that depends on a parameter vector $y := (y_1, \dots, y_d) \in \mathbb{R}^d$. We denote by $U \subset \mathbb{R}^d$ the parameter domain that describe the range of values of y, and assume that there exists a fixed Banach space V, typically a Sobolev space, where the problem (1.1) is well posed for every $y \in U$. We may therefore define the *solution map* from U to V:

$$u : y \mapsto u(y), \tag{1.2}$$

which associates to every parameter $y \in U$, the unique solution $u(y) \in V$ of (1.1).

Parametric PDEs are used to model complex systems in a variety of physical and engineering contexts. Without going into an exhaustive classification of theses contexts, we make the following major distinction:

- **Deterministic modelling:** The parameters y are deterministic inputs of the physical system that can be controlled and monitored by the user. They could for instance be *design or control parameters* in a real or numerically simulated industrial process. A typical application in this context is the optimization of a certain scalar quantity of interest Q that depends on the solution and therefore on the parameters:

$$y \mapsto u(y) \mapsto Q(u(y)). \tag{1.3}$$

For example, consider the steady state heat equation set on a domain D

$$-\text{div}(a\nabla u) = f \quad \text{in} \quad D, \quad u_{|\partial D} = 0, \tag{1.4}$$

with f a given source term and $a = a(y)$ picked from a family $\{a(y) \ : \ y \in U\}$ of thermal conductivity functions. We may use the parameter y in the design of the material with the objective of minimizing the heat flux of the temperature fields $u(y)$ over a portion of the boundary $\Gamma \subset \partial D$. In this case, the scalar quantity of interest is

$$y \mapsto Q(y) = \int_\Gamma \frac{\partial u(y)}{\partial n}(x)dx. \tag{1.5}$$

- **Stochastic modelling:** The parameters y are realizations of random variables which reflect uncertainties in the physical model described by (1.1). For instance, if the equation (1.4) is used to model the diffusion in a porous media which properties are not known exactly, it is then natural to model the diffusion coefficient a as a random field, which as explained further may be described by a sequence $(y_j)_{j\geq 1}$ of scalar random variables. In stochastic modelling, the user is typically interested in the resulting statistical properties of the solution u, which is itself a random field over V. For instance, one may want to compute, if it exists, the average field $\bar{u} := \mathbb{E}[u]$ which is a deterministic function in V, the standard deviation $\mathbb{E}[\|u - \bar{u}\|_V^2]$, the expectation of a scalar quantity of interest $Q = Q(y)$ that depends on the solution similar to the previous deterministic context, or a confidence interval for this quantity.

In addition to the distinction between deterministic and random contexts, the parameters $(y_j)_{j\geq 0}$ may be used to describe very different quantities: the conductivity or diffusion properties of material as in the above mentioned examples, the flux function in a transport problem, a forcing term such as the right hand side in (1.4), the geometry of the physical domain (through a parametrization of the boundary, for instance using control points in computer aided design). It is also possible that several such quantities are simultaneously considered, meaning that y concatenates all parameters used for describing the different quantities.

A significant part of this thesis is devoted to the study of the model problem (1.4) for a particular class of coefficients a. Although simple in formulation, it is relevant for establishing a methodology for the treatment of other classes for parametric PDEs. Here, $D \subset \mathbb{R}^m$ a bounded Lipschitz domain, with m typically equal to 2 or 3, and f in $H^{-1}(D)$. We consider the second order elliptic problem

$$-\text{div}(a(y)\nabla u) = f \quad \text{in} \quad D, \quad u_{|\partial D} = 0, \tag{1.6}$$

where for every $y \in U$, the diffusion function $a(y) \in L^\infty(D)$ has affine dependance on y, according to

$$a(y) := \overline{a} + \sum_{j=1}^{d} y_j \psi_j \tag{1.7}$$

where $\overline{a}$ and the ψ_j are functions in $L^\infty(D)$. In addition, we assume that the problem is *uniformly elliptic* over U, in the sense that there exist $0 < r \leq R < \infty$ such that

$$r \leq a(x,y) \leq R, \quad x \in D,\ y \in U, \tag{1.8}$$

where we have used the notation

$$a(x,y) := a(y)(x) = \overline{a}(x) + \sum_{j=1}^{d} y_j \psi_j(x). \tag{1.9}$$

Under such assumptions, Lax-Milgram theory ensures that the problem (1.6) is well-posed in $V := H_0^1(D)$ for every $y \in U$. The solution map associates to every $y \in U$ a unique solution $u(y) \in V$.

Assuming an affine dependence in y for $a(y)$ is relevant in several contexts. For example if a is piecewise constant over a disjoint partition $D = \cup_{j=1}^{d} D_j$ of the physical domain D, then it is natural to set

$$a(y) = \overline{a} + \sum_{j=1}^{d} y_j \chi_{D_j}, \tag{1.10}$$

where $\overline{a}$ is a constant and χ_{D_j} the indicator function of D_j. More generally the affine form (1.7) is encountered if we truncate the expansion of $a - \overline{a}$, where $\overline{a}$ is a function in x, in a given basis $(\psi_j)_{j\geq 1}$, that is, an expansion of the form

$$a(y) = \overline{a} + \sum_{j=1}^{\infty} y_j \psi_j. \tag{1.11}$$

There are of course many possible choices for such a basis (Fourier series, orthogonal polynomials, wavelets...). In the stochastic context, when a is second order random field, with expectation $\mathbb{E}[a] = \overline{a}$ and continuous covariance function

$$(x,z) \in D \times D \mapsto cov[a](x,z) := \mathbb{E}[(a(x) - \overline{a}(x))(a(z) - \overline{a}(z))], \tag{1.12}$$

a frequently used choice is the *Karhunen-Loève basis*, in other words the orthonormal eigenfunctions of the operator

$$v \mapsto T_a v := \int_D cov[a](\cdot, x) v(x) dx, \qquad v \in L^2(D), \tag{1.13}$$

which is compact, self-adjoint and non-negative on $L^2(D)$. The resulting scalar variable y_j are centered and mutually uncorrelated, i.e. $\mathbb{E}[y_j] = 0$ and $\mathbb{E}[y_i y_j] = \delta_{ij}$ for $i, j \geq 1$, with variance given by the corresponding eigenvalue $\lambda_j > 0$.

Throughout the rest of this thesis, we assume that the parameter domain U has a simple tensor product form, by which we mean that the variables y_j may vary independently in intervals I_j. Such an assumption is quite natural for deterministic problems where these parameters may be tuned independently. For instance in the piecewise constant model (1.10), these intervals could be of the form $I_j = [-\alpha_j, \alpha_j]$ with $0 < \alpha_j < \overline{a}$ for all j. In the stochastic context, using for example the above Karhunen-Loève representation, this assumption is natural if one assumes that the y_j are independent random variables. Note that statistical independence of the component is a stronger property than decorrelation. We may then, upon rescaling the basis functions ψ_j, assume in both the deterministic and stochastic contexts that the parameter domain U is the unit hypercube in d dimension,

$$U := [-1, 1]^d. \tag{1.14}$$

In models where the parameters $y = (y_1, \ldots, y_d)$ correspond to the truncation of an infinite series such as in (1.11), the accuracy is affected by the level d of truncation. In order to reach arbitrarily high accuracy in the numerical approximation of such models, one therefore needs to allow the number of variables d to grow. As explained further, this growth has in principle a severe computational cost expressed by the *curse of dimensionality*. One of the objective of this thesis is to develop numerical methods that are as much as possible immune to the growth of the truncation level d, in the sense that they readily apply to the case where

$$y = (y_j)_{j \geq 1}, \tag{1.15}$$

is infinite dimensional. In this case, the rescaled parameter domain is the infinite dimensional hypercube,

$$U := [-1, 1]^{\mathbb{N}} \tag{1.16}$$

2 Numerical approximation

In both deterministic and stochastic setting, concrete applications may in principle require the evaluation of the solution $u(y)$ for a very large number N of instances of the parameter vector y. Typical examples are the optimisation of a scalar quantity of interest $y \mapsto Q(u(y))$, for instance using Newton's method, or the approximation of the average $\mathbb{E}[Q(u(y))]$ by Monte Carlo methods. Such approaches thus require queries $\{u_i = u(y^i) \ : \ i = 1, \ldots, N\}$ of the solution map (1.2), each of them being approximately executed by a numerical solver which may be computationally expensive in the case of a complex system. It should also be noted that the mentioned approaches are *goal oriented*, in other words, the database of gathered evaluations for a certain task (such as optimization) is unlikely to be used for another task (such as averaging).

In light of the above difficulties, a typical challenge is to simultaneously approximate all the solution $u(y)$ for $y \in U$ to some prescribed accuracy, at reasonable computational cost, which amounts in approximating the solution map $u : y \mapsto u(u)$.

This task is difficult since in contrast to the standard problem of approximating a real-valued function $u : \mathbb{R} \mapsto \mathbb{R}$, the solution map u associated to a parametric PDE (i) is defined on a multidimensional domain $[-1, 1]^d$ where the parametric dimension d can be large, or even infinite, and (ii) takes its value in an infinite dimensional space V, or in a finite yet large dimensional discretization subspace $V_h \subset V$ when using a given numerical solver.

The first item (i) points out the issue of the curse of dimensionality which refers to the exponential blow up of complexity occurring in discretization methods, as the number d of variables grows, even for $\mathbb{R}$-valued functions. Another expression of this phenomenon is the deterioration of approximation rates as d grows, for functions of a given smoothness: for example the accuracy in the L^∞ (or uniform) metric of reconstructing an arbitrary function with continuous derivatives up to order m by piecewise polynomials from h-spaced grid samples is at best of order h^m and therefore, in terms of the number of degrees of freedom n, of asymptotic order $n^{-m/d}$, which is a very poor convergence rate when d is large. A deeper investigation in terms of nonlinear width theory [43, 40, 80] reveals that this poor convergence rate cannot be improved by any other discretization method.

The second item (ii) is concerned with the practical implementation of approximations. The instances $u(y)$ of the solution map or any related quantity, for instance the coefficients of a polynomial approximation in the parametric variable y to this map, can only be computed approximately by space discretization, such as by finite element methods. Therefore, it is crucial to incorporate these considerations on the analysis of the final numerical error. Numerous questions may arise when analyzing discretization errors. For instance, should one use the same discretization space V_h for all instances? is the approximation method of the solution map u robust to discretization errors? etc. We leave aside the space discretization in the remainder of this introduction and focus our attention on the parametric discretization.

We distinguish two approaches for the approximation of the solution map (1.2). An inherent property of both approaches is the separation of the parametric vector y and the physical variable x, space and/or time, in the approximation of u. The first approach consists in building a cheaply computable map

$$y \in U \mapsto u_n(y) := \sum_{i=1}^{n} v_i \phi_i(y) \in V, \tag{2.1}$$

based on a small number n of functions $v_i \in V$ and scalar valued functions ϕ_i from U to $\mathbb{R}$ or $\mathbb{C}$. For example, the v_i could be particular instances of the solution u associated with well chosen values $y^i \in U$ of the parameter vector, that is $v_i = u(y^i)$, and the ϕ_i

could be the Lagrange functions in a given polynomial interpolation process associated to the points $(y^1, \dots, y^n)$. In the stochastic setting, these methods are commonly called *spectral stochastic methods*, see [51, 52, 59].

Depending on the modelling context, deterministic or stochastic, and on the aimed application, one decides in which way the approximation u_n should be close to u. If the objective is capturing the map u everywhere in U to a prescribed accuracy $\varepsilon(n)$, then the error should be considered in the *uniform sense*, i.e.

$$\sup_{y\in U} \|u(y) - u_n(y)\|_V \leq \varepsilon(n). \tag{2.2}$$

In the stochastic context, the quality of approximation is often measured in an *average sense*, such as a mean square error estimate of the form

$$\mathbb{E}[\|u(y) - u_n(y)\|_V^2] := \int_U \|u(y) - u_n(y)\|_V^2 d\varrho(y) \leq \varepsilon^2(n), \tag{2.3}$$

where ϱ is the joint probability distribution of random vector y. Note that the first estimate implies the second estimate.

The second approach consists in searching for a subspace E_n of V of low dimension n that could serve for simultaneous approximation of all solutions, for example using the Galerkin method. This means that we aim to approximate the *solution manifold*

$$\mathcal{M} := \Big\{u(y) : y \in [-1,1]^d\Big\} \subset V, \tag{2.4}$$

by the linear space E_n. Once again, we may search for error estimates in a uniform or mean square sense between u and its best approximation $y \mapsto u_{E_n}(y) := \operatorname{argmin}_{v\in E_n}\|u(y) - v\|_V$, which is computed via the orthogonal projection of every $u(y)$ onto E_n in the case where V is a Hilbert space.

In the case of uniform estimates, the optimal choice for E_n, if it exists, corresponds to the space that achieves the Kolmogorov n-width of the solution manifold in V, that is

$$d_n(\mathcal{M})_V := \inf_{\dim(E)\leq n} \sigma_E(\mathcal{M}), \qquad \sigma_E(\mathcal{M}) := \sup_{w\in\mathcal{M}} \inf_{v\in E} \|w - v\|_V. \tag{2.5}$$

In the stochastic setting, one usually substract the average field $\overline{u} = \mathbb{E}(u)$ to u and searches for the space E_n that minimizes the least-square error between $\tilde{u} = u - \overline{u}$ and its best approximation $\tilde{u}_E$, that is,

$$\mathbb{E}(\|\tilde{u} - \tilde{u}_E\|_V^2), \tag{2.6}$$

among all n-dimensional spaces E. The optimal choice is related to Hilbert-Karhunen-Loève expansion

$$\tilde{u} = \sum_{j=1}^{\infty} \sqrt{\lambda_i} v_j U_j \tag{2.7}$$

where $(\lambda_j, v_j)_{j\geq 1}$ the family of decreasing eigenvalues and orthonormonal eigenvectors associated to the covariance operator of u in V, see [45] for more details, and

$$U_j := \frac{1}{\sqrt{\lambda_i}}\langle \tilde{u}, v_j\rangle_V, \tag{2.8}$$

are centred and mutually uncorrelated random variables with variance 1. The optimal space E_n is then spanned by $\{v_1, \ldots, v_n\}$.

In both cases, the optimal spaces are out of reach from a computational point of view, and one therefore needs to rely on sub-optimal yet more easily computable choices. For least-squares estimates, one may approximate $\overline{u}$ and the covariance kernel $cov[u]$ using the knowledge of u on a coarse descritisation of V, see [45], or by computing $\overline{u}$ and $cov[u]$ without any knowledge on u, see [65] for the problem (1.6). However, the computation of the Hilbert-Karhunen-Loève expansion eventually amounts to the resolution of a generalized eigenvalue problem which can be numerically costly. For uniform estimates, a popular strategy in this direction is the *reduced basis* method [16, 64, 63]. In this strategy, one first acquires during a possibly expensive *off-line* processing stage n elements $v_j := u(y^j)$ of the manifold $\mathcal{M}$, then in the *online* stage, approximate for any parameter query y the solution $u(y)$ in the space $F_n := \mathrm{span}\{v_1, \ldots, v_n\}$ by a Galerkin scheme. The spaces F_n are generally sub-optimal compared to the n-width spaces E_n that achieve the infimum in (2.5). However, it was proved in [12] and [42] that a certain *greedy* selection of the instances v_i in the off-line stages produces spaces that are rate-optimal in the following sense: for $\alpha, c > 0$, it is proved that

$$\sup_{n>0} n^\alpha \sigma_{F_n}(\mathcal{M}) \leq C_\alpha \sup_{n>0} n^\alpha d_n(\mathcal{M}), \tag{2.9}$$

and

$$\sup_{n>0} e^{c_\alpha c n^\alpha} \sigma_{F_n}(\mathcal{M}) \leq C_\alpha \sup_{n>0} e^{cn^\alpha} d_n(\mathcal{M}), \tag{2.10}$$

where $C_\alpha, c_\alpha > 0$ are constants depending on α.

It should be well understood that the two above approaches are connected. On the one hand, if u_n is an approximation to u of the form (2.1) in the uniform sense to accuracy $\varepsilon(n)$, then by introducing the space $F_n := \mathrm{span}\{v_1, \ldots, v_n\}$, we obviously have

$$\sigma_{F_n}(\mathcal{M}) = \sup_{y\in U} \inf_{v\in F_n} \|u(y) - v\|_V \leq \sup_{y\in U} \|u(y) - u_n(y)\|_V \leq \varepsilon(n). \tag{2.11}$$

Therefore the linear space F_n approximates the manifold $\mathcal{M}$ in the uniform sense with accuracy $\varepsilon(n)$, implying an estimation of the Kolmogorov n-width by

$$d_n(\mathcal{M}) \leq \sigma_{F_n}(\mathcal{M}) \leq \varepsilon(n), \tag{2.12}$$

which may in turn be used to study the convergence of the reduced basis method, in view of (2.9) and (2.10). One the other hand, if $E_n := \mathrm{span}\{v_1, \ldots, v_n\}$ is the space

achieving the Kolmogorov n-width, when it is attained,then we may write for every $y \in U$

$$u_{E_n}(y) = \sum_{i=1}^{n} v_i \phi_i(y), \tag{2.13}$$

where $\phi_i(y)$ are the coordinate of $u_{E_n}(y)$ associated with the v_i. Obviously u_{E_n} is of the form (2.1) and

$$d_n(\mathcal{M}) = \sigma_{E_n}(\mathcal{M}) = \sup_{y \in U} \|u(y) - u_{E_n}(y)\|_V \tag{2.14}$$

which shows that u_{E_n} is the best approximation of the form (2.1) in the uniform sense.

In practice, there is however an essential difference between the two approaches. In the first approach, we first choose n scalar valued functions $\phi_1, \ldots, \phi_n$ in a family of functions, for example the family of Legendre polynomials, then we compute the functions $v_1, \ldots, v_n \in V$. In the second approach, one rather strives to identify n "good" functions $v_1, \ldots, v_n$ in V and then for every given y compute the values $\phi_j(y)$ by the Galerkin projection method in the space spanned by these functions. Once the functions $v_1, \ldots, v_n$ are computed, the first approach has the advantage that the computation of $u_n(y)$ is immediate by the linear combination (2.1) while a system inversion is required, for every $y \in U$, in the second approach.

This thesis is only concerned with the first approach, namely the construction of cheaply computable approximations of the form (2.1) to the solution map, with the functions $\phi_i(y)$ being particular multivariate polynomials in the y variable. Another important distinction between numerical methods for parametric PDE's is through the two following classes, both of them being considered in this thesis:

- ***Non-intrusive methods*** rely only on repeated (approximate) queries of the solution map u, obtained by an existing deterministic numerical solver. Typically, this solver is an industrial simulation code, considered as a black-box, it can associate to every parameter vector $y \in U$ an output $u^\varepsilon(y) \in V$ approximating $u(y)$ to any desired accuracy $\varepsilon > 0$. One example of such methods is the Monte-Carlo approximation of the average field $\overline{u}$ by the empirical mean

$$\overline{u}_n := \frac{1}{n} \sum_{i=1}^{n} u(y^i), \tag{2.15}$$

  where the y^i are independent realizations of the random parameter vector y. Another example is the approximation of the solution map by interpolation at chosen points $y^1, \ldots, y^n \in U$. Non-intrusive methods are convenient in that they may be thought as a post-processing on top of existing numerical solvers. However the dependence on the solver, which itself could be computationally expansive, might be a serious limitation. In addition, when the solver is given as a black box with no precise

informations on the underlying PDE model (1.1), one has no theoretical guarantee that a given non-intrusive method is suited for the numerical purpose or that it even produces converging approximations.

- ***Intrusive methods*** in contrast exploit the features of specific classes of parametric and stochastic PDEs. By specific, we mean any additional information (data or assumption) on the equation (1.1) governing the physical system. For example, the distribution of the random vector y in the stochastic context, the category of the operator $\mathcal{D}$, its dependence on the parameters y, etc. The knowledge of such specific features allows the design of methods that are well adapted to the problem, hence hopefully outperforms non intrusive methods in precision and speed. For example, knowing the exact model allows to use the Galerkin method for the parametric discretization using for example polynomials in the y variable. Note that the previously described reduced basis method is non-intrusive in the offline stage that computes instances of solutions, but intrusive in the online stage that uses these instances to generate a particular trial space for the Galerkin method.

3 Polynomial approximations of the solution map

As already discussed, there exists n-term approximations with the separable form (2.1) that are optimal, either in the uniform or least-square sense, however they are not easily computable. In addition, the corresponding functions ϕ_j may be quite complex since they depend on all the variables y_j. One approach is to search for further separation of variables through approximations of the form

$$u_n(y) = \sum_{j=1}^{n} v_j \prod_{i\geq 1} \phi_{j,i}(y_i). \tag{3.1}$$

The problem of an optimal choice of the factors is not well posed. For illustrative purposes, we assume that u is real valued and that the parameter domain U is discrete and finite of the form $\{t_1, \ldots, t_k\}^d$. In this case $V := R$ and $u \in V^{k\times\cdots\times k}$ is an order-d tensor and approximations of the form (3.1) are rank-n approximations of u, which are of the form

$$u_n = \sum_{j=1}^{n} v_j \; \Phi_{j,1} \otimes \ldots \otimes \Phi_{j,d}, \tag{3.2}$$

where $\Phi_{j,i} := (\phi_{j,i}(t_1), \ldots, \phi_{j,i}(t_k))^t$. It is well known that the problem of finding the best n-rank, for $n \geq 2$, is in general ill-posed and for all norms, except for $d = 2$ where it is completely resolved for the Hilbert-Schmidt and spectral norms, thanks to Eckart-Young Theorem, see [38].

Alternate approaches consist in picking the functions $\phi_{j,i}$ among a predefined family. For instance, we may impose that the functions $\phi_{j,i}$ are picked among polynomials,

piecewise polynomials, trigonometric functions, etc. Here, we consider polynomials, in which case approximation of the form (3.1) are multivariate polynomials over U with coefficients in V.

Approximations by tensorized polynomials may be constructed using various methods, for example Taylor series, Legendre series, Galerkin projection, polynomial interpolation, least squares, sparse grids, etc, each of which favored depending on the approximation purpose and the amount of knowledge on the parametric PDE at hands. In particular, among the previously cited methods, only the last three are non-intrusive. For every method, the polynomials $\phi_{j,i}$ have a predefined form. We introduce next a unified notation that we use to describe various polynomial methods introduced in recent years.

We consider a family of univariate polynomials

$$\mathcal{P} := (P_j)_{j\geq 0}, \tag{3.3}$$

with P_0 constant equal to 1 and P_j of degree exactly j, so that $\{P_0, \dots, P_k\}$ is a basis of $\mathbb{P}_k$ the space of polynomials of degree at most k. We denote by $\mathcal{F}$ the set $\mathbb{N}^d$ of multi-indices of length d. We consider the family $(P_\nu)_{\nu\in\mathcal{F}}$ of tensorized multivariate polynomials defined by

$$P_\nu(y) = \prod_{j=1}^{d} P_{\nu_j}(y_y), \quad y = (y_1, \dots, y_d), \qquad \nu \in \mathcal{F}. \tag{3.4}$$

The polynomial approximations that we consider have the form

$$u_\Lambda := \sum_{\nu\in\Lambda} v_\nu \, P_\nu, \tag{3.5}$$

where $\Lambda \subset \mathcal{F}$ is a set of finite cardinality and $\{v_\nu\}_{\nu\in\Lambda_n}$ are elements in V. The polynomials u_Λ thus belong to the space of V-valued polynomials over U,

$$\mathbb{V}_\Lambda := V \otimes \mathbb{P}_\Lambda \quad \text{where} \quad \mathbb{P}_\Lambda = \mathbb{P}_\Lambda(\mathcal{P}) := \operatorname{span}\{P_\nu : \nu \in \Lambda\}. \tag{3.6}$$

It is important to notice that the space $\mathbb{V}_\Lambda$ depends in principle both on the set Λ and on the chosen family $\mathcal{P}$. For the family $\mathcal{P}$, we may typically consider the monomials,

$$P_j(t) = t^j, \tag{3.7}$$

or the Legendre polynomials obtained by Gram-Schmidt orthogonalization of the latter in $L^2([-1,1], \frac{dt}{2})$. However, if Λ has the property that

$$\nu \in \Lambda \ \text{ and } \ \mu \leq \nu \Rightarrow \mu \in \Lambda, \tag{3.8}$$

where $\mu \leq \nu$ means that $\mu_i \leq \nu_i$ for all $i = 1, \ldots, d$, then it is readily seen that $\mathbb{V}_\Lambda$ is independent of $\mathcal{P}$, in which case

$$\mathbb{V}_\Lambda := V \otimes \operatorname{span}\{y \mapsto y^\nu : \nu \in \Lambda\} \quad \text{where} \quad y^\nu = \prod_{j=1}^{d} y_j^{\nu_j}. \tag{3.9}$$

Sets Λ having this property are called *lower sets* or *downward closed sets*, and play an important role in this thesis.

For a given parametric or stochastic PDE (1.1), two basic questions are the following:

(i) How to identify a set Λ of small cardinality such that the solution map is well approximated in $\mathbb{V}_\Lambda$?

(ii) How to practically compute an approximation of the solution map in $\mathbb{V}_\Lambda$?

In the case of the model elliptic problem (1.6) with affine dependance as in (1.7) and uniform ellipticity assumption (1.8), these questions have been addressed in recent years by many different approaches [2, 5, 6, 4, 68, 70, 69, 7, 8, 74, 34, 33, 22]. Practical algorithms of polynomial approximation considered in these works also incorporate discretization in the space variable, taking into account the fact that the instances $u(y) \in V$ or any other coefficient in V associated with u can only be approximated to a given accuracy. For instance, for model (1.6), this can be done through discretization by the finite element method where functions $v \in V = H_0^1(D)$ are approximated in continuous, piecewise linear finite element spaces $V_h \subset V$ on regular quasi-uniform simplicial partitions of D of meshwidth h, see e.g. [28] for a general introduction. The FEM error is then incorporated in the analysis of the overall approximation error. At this stage, this is irrelevant for our exposition and we describe the polynomial methods in a semi-discrete setting, that is, only the discretization in the parametric variable is considered.

The approach proposed in [2] is based on Neumann series applied to the operator formulation of the problem

$$A(y)u = f, \tag{3.10}$$

where for every $y \in U$, $A(y)$ is the differential operator from V to V^* defined by $A(y)v = -\operatorname{div}(a(y)\nabla v)$. In view of the (1.7), the operator can be decomposed into $A = A_0 + \Psi$ where A_0 given by $A_0 v = -\operatorname{div}(\overline{a}\nabla v)$ does not depends in y and $\Psi(y)v = -\operatorname{div}((a(y) - \overline{a})\nabla v)$. Under the uniform ellipticity assumption (1.8), we write

$$u(y) = A(y)^{-1} f = (Id + A_0^{-1}\Psi(y))^{-1} g, \qquad g := A_0^{-1} f, \tag{3.11}$$

with $\|A_0^{-1}\Psi\|_{V\to V} \leq \xi = 1 - \frac{r}{R} < 1$. This allows to apply the Neumann series expansion and obtain the exponential bound

$$\sup_{y\in U} \left\| u(y) - \sum_{j=0}^{k} (-1)^j (A_0^{-1}\Psi(y))^j g \right\|_V \lesssim \xi^k. \tag{3.12}$$

Since the operator $\Psi(y)$ has an affine dependence y, then the polynomial in the previous approximation belongs to $\mathbb{V}_{\mathcal{S}_k}$ where $\mathcal{S}_k$ is the simplex

$$\mathcal{S}_k := \Big\{\nu \in \mathbb{N}^d : |\nu| := \sum_{j=1}^{d} \nu_j \leq k\Big\}. \tag{3.13}$$

Obviously $\mathcal{S}_k$ is a lower set and $\mathbb{P}_{\mathcal{S}_k}$ is the space of d-variate polynomials of total degree k. The polynomial approximation converges exponentially fast toward u in the uniform sense, however the polynomials space has dimension $\binom{k+d}{k}$ which grows fast with d and k.

In the subsequent works [5, 6], approximations of u in the mean square sense are constructed by Galerkin projection on predefined piecewise polynomial and polynomial spaces. In the stochastic setting, denoting ϱ the joint probability distribution of the random vector y, the map u can be defined as the unique function in the Bochner space $\mathcal{V}_2 := L^2(U, V, d\varrho)$, solution of the variational problem

$$\int_U \int_D a(y)\nabla u(y)\nabla w(y) d\varrho(y) = \int_U \int_D f w(y) d\varrho(y), \qquad w \in \mathcal{V}_2, \tag{3.14}$$

and its Galerkin approximation in $\mathbb{V}_\Lambda$ is the unique $u_\Lambda \in \mathbb{V}_\Lambda$ such that

$$\int_U \int_D a(y)\nabla u_\Lambda(y)\nabla w(y) d\varrho(y) = \int_U \int_D f w(y) d\varrho(y), \qquad w \in \mathbb{V}_\Lambda. \tag{3.15}$$

The polynomial spaces considered in [5, 6] are of type $\mathbb{V}_{\mathcal{B}_\mu}$ where for $\mu \in \mathbb{N}^d$, $\mathcal{B}_\mu$ is the rectangular block

$$\mathcal{B}_\mu := \Big\{\nu \in \mathbb{N}^d : \nu \leq \mu\Big\}. \tag{3.16}$$

We note that $\mathcal{B}_\mu$ is a lower set and that $\mathbb{P}_{\mathcal{B}_\mu}$ is the space of d-variate polynomials of degree at most μ_j in variable y_j. The authors assume that ϱ is the product of the uniform product measure $\hat{\varrho} := \otimes_{j=1}^d \frac{dy_j}{2}$ by a bounded function. This, combined with the optimality of the Galerkin projection $u_{\mathcal{B}_\mu}$ and the uniform ellipticity assumption implies

$$\mathbb{E}\Big[\|u(y) - u_{\mathcal{B}_\mu}(y)\|_V^2\Big] \lesssim \int_U \|u(y) - \sum_{\nu \in \mathcal{B}_\mu} v_\nu L_\nu(y)\|_V^2 d\hat{\varrho}(y) = \sum_{\nu \notin \mathcal{B}_\mu} \|v_\nu\|_V^2, \tag{3.17}$$

where $(L_\nu)_{\nu \in \mathcal{F}}$ are the tensorized Legendre polynomials, orthonormal with respect to $\hat{\rho}$, and v_ν are the associated coefficients. Deriving estimates on the V-norm of Legendre coefficients $\|v_\nu\|_V$ through the study of the partial derivative of u and using the product structure of the polynomials space $\mathbb{P}_{\mathcal{B}_\mu}$, the authors show that $u_{\mathcal{B}_\mu}(y)$ converge toward u with a bound that is roughly of the form

$$\mathbb{E}\Big[\|u(y) - u_{\mathcal{B}_\mu}(y)\|_V^2\Big] \lesssim \sum_{j=1}^{d} \Big(1 + \frac{c}{\|\psi_j\|_{L^\infty(D)}}\Big)^{-(\mu_j+1)}, \tag{3.18}$$

with c a fixed constant. Note however that if $\mu_j \geq 1$ for every j, then the dimension of the polynomial space $\mathbb{P}_{\mathcal{B}_\mu}$ exceeds 2^d, which reflects the curse of dimensionality.

In the case where ϱ is a product measure, which is equivalent to the independence of the random variables y_j, the authors propose to use tensorized *double orthogonal polynomials* in order to decouple the Galerkin system and compute rapidly the Galerkin projection. Without loss of generality, suppose that $\varrho := \otimes_{j=1}^d \frac{dy_j}{2}$ and denote $(L_j)_{j\geq 1}$ the Legendre polynomials orthonormal in $L^2([-1,1], \frac{dt}{2})$. Since $\mathbb{P}_{\mathcal{B}_\mu}$ has a product structure, one has

$$\mathbb{P}_{\mathcal{B}_\mu} = \otimes_{j=1}^d \text{span}\left\{ l_k^{\mu_j+1} \ : \ k = 0, \dots, \mu_j \right\} = \text{span}\{ l_\nu^\mu : \nu \leq \mu \} \tag{3.19}$$

where

$$l_k^n := \frac{L_n}{(t - t_k^n) L_n'(t_k^n)} \quad \text{and} \quad l_\nu^\mu := \otimes_{j=1}^d l_{\nu_j}^{\mu^j+1}, \qquad \nu \leq \mu, \tag{3.20}$$

and for every $n \geq 1$, $t_0^n, \dots, t_{n-1}^n$ are the n simple roots of the Legendre polynomial L_n. It is easy to show, using elementary orthogonality arguments, that for every $n \geq 1$ and $0 \leq i, j \leq n-1$

$$\int_{-1}^1 l_i^n(t) l_j^n(t) \frac{dt}{2} = \beta_i^n \delta_{i,j}, \quad \int_{-1}^1 t \ l_i^n(t) l_j^n(t) \frac{dt}{2} = t_i^n \beta_i^n \delta_{i,j}, \quad \text{where} \ \ \beta_i^n = \frac{1}{2} \int_{-1}^1 (l_i^n)^2 dt. \tag{3.21}$$

Since a depends linearly on y as in (1.7), then by formulating the Galerkin system using the polynomials $\{l_\nu^\mu\}_{\nu \in \mathcal{B}_\mu}$, it can be easily shown that the corresponding coordinates $u_{\mathcal{B}_\mu,\nu}$ of the Galerkin projection $u_{\mathcal{B}_\mu}$ in the previous basis, are the unique solutions in $H_0^1(D)$ of the following variational problem s

$$\beta_\nu^\mu \int_D (\overline{a} + \sum_{j=1}^d t_{\nu_j}^{\mu_j+1} \psi_j) \nabla u_{\mathcal{B}_\mu,\nu} \nabla w = \int_U \int_D f w l_\nu^\mu d\varrho, \qquad w \in H_0^1(D), \tag{3.22}$$

where $\beta_\nu^\mu = \prod_{j=1}^d \beta_{\nu_j}^{\mu_j+1}$. The computation of Galerkin projection amounts then to solving $\prod_{j=1}^d (1+\mu_j)$ deterministic boundary problems equivalent in cost to computing ane instance of the solution map u. In addition, we should note the following remark that was not mentioned in [5]. If f does not depend on y, then the right side in (3.22) is a product of two integrals and since the solutions $u(y)$ of (1.6) satisfy the following variational problems

$$\int_D (\overline{a} + \sum_{j=1}^d y_j \psi_j) \nabla u(y) \nabla w = \int_D f w, \qquad w \in H_0^1(D), \tag{3.23}$$

then by denoting $t_\nu^\mu := (t_{\nu_1}^{\mu_1+1}, \dots, t_{\nu_d}^{\mu_d+1}) \in U$, one gets that

$$u_{\mathcal{B}_\mu,\nu} = \frac{w_\nu^\mu}{\beta_\nu^\mu} u(t_\nu^\mu), \qquad w_\nu^\mu := \int_U l_\nu^\mu(y) d\varrho(y) = \prod_{j=1}^d \int_U l_{\nu_j}^{\mu_j+1}(t) \frac{dt}{2} \tag{3.24}$$

We note that w_ν^μ is the Gauss weights associated with the multidimensional abscissas t_ν^μ for quadratures in the grids of point $\{t_\nu : \nu \in \mathcal{B}_\mu\}$.

In a later work [4], the authors propose to compute an approximation of u in the space $\mathbb{V}_{\mathcal{B}_\mu}$ directly by first collocating the variational problem (3.23) satisfied by the functions $u(y) \in V$, in the tensorized grid

$$\Gamma_{\mathcal{B}_\mu} := \{t_\nu^\mu : \nu \leq \mu\} = \otimes_{j=1}^d \{t_\nu^\mu : \nu \leq \mu\}, \tag{3.25}$$

then constructing the approximation by interpolation. Thanks to the product structure of $\mathbb{P}_{\mathcal{B}_\mu}$, it is easy to see that $\{l_\nu^\mu\}_{\nu \in \mathcal{B}_\mu}$ are the Lagrange polynomials associated with the grid $\Gamma_{\mathcal{B}_\mu}$ in the space $\mathbb{P}_{\mathcal{B}_\mu}$, so that the interpolation operator is given by

$$\mathcal{I}_\mu u := \sum_{\nu \leq \mu} u(t_\nu^\mu) l_\nu^\mu. \tag{3.26}$$

One way to analyze the stability of the interpolation operator is to use the double orthogonality property in (3.21). More precisely, with $\|u\|_{\mathcal{V}_\infty} := \sup_{y \in U} \|u(y)\|_V$, we have

$$\mathbb{E}[\|\mathcal{I}_\mu u(y)\|_V^2] = \sum_{\nu \leq \mu} \|u(t_\nu^\mu)\|_V^2 \int_U (l_\nu^\mu(y))^2 d\varrho(y) \leq \|u\|_{\mathcal{V}_\infty}^2 \sum_{\nu \leq \mu} \int_U (l_\nu^\mu(y))^2 d\varrho(y) = \|u\|_{\mathcal{V}_\infty}^2, \tag{3.27}$$

where the last identity follows from the orthogonality property (3.21) and the fact that $\sum_{\nu \leq \mu} l_\nu^\mu(y) = 1$. Therefore, one has

$$\mathbb{E}[\|u - \mathcal{I}_\mu u\|_V^2] \leq 2 \inf_{v \in \mathbb{V}_{\mathcal{B}_\mu}} \sup_{y \in U} \|u(y) - v(y)\|_V^2. \tag{3.28}$$

Investigating more thoroughly the growth of partial derivatives of u as in [5], the authors show that the map u admits an holomorphic extension in the complex domain and use it to show that the right side in the last inequality satisfy a similar bound as the L^2 error in (3.18).

In [70, 69], polynomial approximations are constructed using collocation methods in polynomials spaces which are not of the above tensor product type, following the *sparse grids* approach, originally due to Smolyak [76] and investigated in many works, among others [50, 71, 9, 81]. In [70], the polynomial approximations are considered in *isotropic* spaces $\mathbb{V}_{m(\mathcal{S}_k)}$, where m is any given increasing function from $\mathbb{N}$ to $\mathbb{N}$ satisfying $m(0) = 0$ and the convention $m(-1) = -1$, $\mathcal{S}_k$ is the simplex in (3.13) and the notation $m(\mathcal{S}_k)$ stands for

$$m(\mathcal{S}_k) := \bigcup_{i \in \mathcal{S}_k} B_{m(i)} \quad \text{with} \quad B_{m(i)} := \Big\{\nu \in \mathbb{N}^d : m(i_j - 1) < \nu_j \leq m(i_j)\Big\}. \tag{3.29}$$

Let us note that $m(\mathcal{S}_k)$ is always a lower set and that it coincides with $\mathcal{S}_k$ when m is the identity. The polynomial approximation is given by Smolyak formula

$$\mathcal{I}_{m(\mathcal{S}_k)} u = \sum_{i \in \mathcal{S}_k} \otimes_{j=1}^d (I_{m(i_j)} - I_{m(i_j - 1)}) u, \tag{3.30}$$

where we have $I_{-1} = 0$, and for every $l \geq 0$, $I_{m(l)}$ is the Lagrange interpolation operator associated with $m(l)+1$ disjoint points $\{r_0, \ldots, r_{m(l)}\}$ in $[-1, 1]$. When the interpolation points of the operators $I_{m(0)}, I_{m(1)}, \ldots$ are the nested sections of an infinite sequence $r_0, r_1, r_0 \ldots$ of mutually distinct points, the operator $\mathcal{I}_{m(\mathcal{S}_k)}$ is an interpolation operator associated with the space $\mathbb{V}_{m(\mathcal{S}_k)}$ and the isotropic sparse grid of points

$$\Gamma_{\mathcal{S}_k} := \Big\{ r_\nu := (r_{\nu_1}, \ldots, r_{\nu_d}) : \nu \in m(\mathcal{S}_k) \Big\}, \tag{3.31}$$

see [7, 9, 26]. A typical challenge is the optimization of the trade-off between the growth of m which dictates the overall number of collocations and a good choice of collocation points which determines the quality of the approximation. In the nested case, the quality of the approximation can be studied through the stability of the interpolation operator $\mathcal{I}_{m(\mathcal{S}_k)}$, measured by its Lebesgue constant. In [70], the classical choice of nested Clenshaw-Curtis points is studied, more precisely, the choice of m and the nested sections of interpolation point associated with every $I_{m(l)}$ is given by

$$m(l) = 2^l, \quad l \geq 1, \quad \text{and} \quad \{r_0, \ldots, r_{m(l)}\} = \Big\{ \cos\Big(\frac{j}{2^l}\pi\Big) : j = 0, \ldots, m(l) \Big\}. \tag{3.32}$$

The individuals operators $I_{m(l)}$ are known to have Lebesgue constants which grow like $\log(2^l)$. Convergence bounds in the uniform sense are obtaind by combining this with the analyticity results proved in [5].

The previous polynomials spaces $\mathbb{V}_{\mathcal{S}_k}$ are isotropic, and accordingly the approximations $\mathcal{I}_{m(\mathcal{S}_k)}u$ are also isotropic, in the sense that the variables y_j play symmetric roles. For highly anisotropic problems, for example when the functions ψ_j in (1.7) have norms $\|\psi_j\|_{L^\infty(D)}$ that strongly vary with j, the solution u inevitably inherits a highly anisotropic dependence on the individual variables y_j. This should then be reflected in the polynomial approximation. In [69], the authors treat this by considering anisotropic version of the space $\mathbb{V}_{\mathcal{S}_k}$ and accordingly anisotropic versions of $\mathbb{V}_{m(\mathcal{S}_k)}$. These versions are characterized by parameters $\alpha := (\alpha_1, \ldots, \alpha_d) \in {\mathbb{R}_+^*}^d$ according to

$$\mathcal{S}_{k,\alpha} := \Big\{ \nu \in \mathbb{N}^d : \langle \nu, \alpha \rangle := \sum_{j=1}^{d} \nu_j \alpha_j \leq k \Big\}, \qquad m(\mathcal{S}_{k,\alpha}) := \bigcup_{i \in \mathcal{S}_{k,\alpha}} B_{m(i)}. \tag{3.33}$$

We note again that such sets are lower sets. The parameter α should reflect the anisotropy of the problem: the smaller is the dependance on the variable y_j, the larger is the value of α_j. In the isotropic case, α has all entries equal to 1. The approximations considered in [69] are constructed using the same Smolyak formula (3.30) with now $i \in \mathcal{S}_{k,\alpha}$. As in the isotropic setting, in the case of points netedness, the approximation operator is an interpolation operator associated with the space $\mathbb{V}_{m(\mathcal{S}_{k,\alpha})}$ and the anisotropic sparse grid of points

$$\Gamma_{\mathcal{S}_{k,\alpha}} := \Big\{ r_\nu := (r_{\nu_1}, \ldots, r_{\nu_d}) : \nu \in m(\mathcal{S}_{k,\alpha}) \Big\}, \tag{3.34}$$

see [7, 26]. Similar to [70], the convergence analysis is based on the stability of the interpolation operator combined with the analyticity results in [5]. It is in particular shown that with the choice of Clenshaw Curtis points, there exists an optimal choice of α, depending on the radii of analyticity of u in each variable y_j, for which the error bound is minimal. Sparse grid collocation methods are extended to more general anisotropic polynomial spaces in [7, 8], for which the sets Λ are adaptively constructed and optimized using the so-called knapsack algorithm.

All the above described strategies provide computable polynomial approximations that converge toward the solution map u in both the uniform and/or the mean square sense. In order to compare their computational efficiency, an appropriate benchmark consists in analyzing the error decay rate as a function of the overall computational cost. For the methods presented so far, the cost of approximation in a given space $\mathbb{V}_\Lambda$ is essentially dominated by the cost of $\#(\Lambda)$ evaluations of u. It is therefore relevant to study the error rates of each method as a function of $n := \#(\Lambda)$. This is, alas, not straightforward since the error rates are for almost all the strategies in form of sum of error contributions. However, a detailed inspection shows that for each method, the rates is not better than

$$\xi^{\lambda(\Lambda)}, \qquad \lambda(\Lambda) := \max_{\nu\in\Lambda}\{\nu_{\max} : \nu_{\max} := \max(\nu_1, \ldots, \nu_d)\}, \tag{3.35}$$

where ξ is a given number in $]0, 1[$ independent of d. We note that $\lambda(\Lambda)$ is the maximal degree attained in at least one variable of the polynomials in $\mathbb{P}_\Lambda$. For isotropic approximations, the cardinalities of the considered sets are $\#(\mathcal{B}_\mu) = (1 + \mu_1)^d$ with $\mu := (\mu_1, \ldots, \mu_1)$, $\#(\mathcal{S}_k) = \binom{k+d}{k}$ and with m the doubling rules in (3.32), we have if $d \geq k$

$$\#(m(\mathcal{S}_k)) = \sum_{i\in\mathcal{S}_k} \#(B_{m(i)}) \geq \sum_{i\in\mathcal{S}_k\cap\{0,1\}^d} 2^{|i|} = \sum_{j=0}^{k} k\binom{d}{j}2^i \geq d2^k. \tag{3.36}$$

Since $\lambda(\mathcal{B}_\mu) = 1 + \mu_1$, $\lambda(\mathcal{S}_k) = k$ and $\lambda(m(\mathcal{S}_k)) = 2^k$, then it is readily verified that

$$\lambda(\mathcal{B}_\mu) \leq (\#(\mathcal{B}_\mu))^{\frac{1}{d}}, \qquad \lambda(\mathcal{S}_k) \leq \frac{k!}{d^k}\#(\mathcal{S}_k), \qquad \lambda(m(\mathcal{S}_k)) \leq \frac{1}{d}\#(m(\mathcal{S}_k)), \tag{3.37}$$

so that in view of (3.35), the rates of isotropic approximations deteriorate with the growth of the dimension d and are then not immune to the curse of dimensionality. The approximation in anisotropic spaces can in some measure overcomes this limitation since the dimension d can be reduced by activating only few variables y_j in the approximation. For instance, setting $\mu_j = 0$ when working with $\mathcal{B}_\mu$ or $\alpha_j >> 1$ when working with $\mathcal{S}_{k,\alpha}$ and $m(\mathcal{S}_{k,\alpha})$ yields the inactivity of the variable y_j in the approximation. However, due to their coupled rigid shape, these anisotropic sets double at least in cardinality, whenever a new coordinate y_j is activated, while only the j^{th} contribution of the error decreases. We note also that for sparse grids, the multiplicative constants in the error bounds depend on d.

In view of the previous discussion, the challenge consists in designing polynomial methods with convergence rates that are robust to the parametric dimension d. An equivalent challenge is to design polynomial methods with provable convergence rate in the infinite dimensional setting $d = \infty$, for example algebraic rates, that is, error bounds of the form Cn^{-s}, where $n = \#(\Lambda)$. Of course, this objective can not be achieved if all the variables y_j weigh equally in the solution u for the reasons explained above. On the other hand, the convergence of an expansion of the type (1.11) enforces some decay of ψ_j making the variables y_j less active as $j \to +\infty$.

In [34, 33], a new paradigm is introduced for treating model (1.6) in the case $d = \infty$ following the previously discussed prescriptions. The following result is proved: if a is as in (1.11) with the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$ belonging to $\ell^p(\mathbb{N})$ for some $p < 1$, then Taylor series

$$\sum_{\nu\in\mathcal{F}} t_\nu y^\nu \quad \text{with} \quad t_\nu := \frac{\partial_\nu u(0)}{\nu!} \in V, \tag{3.38}$$

truncated to well chosen sets $(\Lambda_n)_{n\geq 1}$, with $\#(\Lambda_n) = n$, converge to u in the uniform sense with at least the algebraic convergence rate n^{-s} up to a multiplicative constant that only depends on $\|(\psi_j)\|_{\ell^p(\mathbb{N})}$, and

$$s := \frac{1}{p} - 1. \tag{3.39}$$

The same result is also proved with Legendre series

$$\sum_{\nu\in\mathcal{F}} v_\nu L_\nu \quad \text{with} \quad v_\nu := \int_U u(y)L_\nu(y)d\varrho(y) \in V, \tag{3.40}$$

in the mean square sense, with the improved convergence rate n^{-s^*}, where

$$s^* := \frac{1}{p} - \frac{1}{2}. \tag{3.41}$$

In other words, algebraic rates that are robust to the dimension d are established, under a certain decay assumption on the functions ψ_j as $j \to +\infty$ that reflects the anisotropy of the solution map.

In order to establish the previous results, a critical idea is to define the sets Λ_n as the indices associated to the n largest terms of the sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ of Taylor and Legendre coefficients. This truncation is a form of nonlinear approximation of the solution map, sometimes called best n-term approximation, see [40]. The sets Λ_n might be quite different from the previously descibed isotropic or anisotropic sets.

The exact knowledge of the best n-term sets is ususally out of reach, and therefore the above results should only be thought as theoretical approximation results. In particular, they do not obviously yield a computational strategy. It should also be

remarked that the proof of the above results strongly exploits the linear elliptic nature of the model (1.6) and the affine parametric dependence in (1.11). These limitations motivate the following natural questions:

(i) Can we build index sets $(\Lambda_n)_{n\geq 1}$ by adaptive or non-adaptive computational strategies, such that the convergence rate for polynomial approximations of the solution map in the space $\mathbb{V}_{\Lambda_n}$ that is similar to that provided by the best $\#(\Lambda_n)$-term approximation ?

(ii) Given such index sets, can we define the associated polynomial approximation of the solution map by simple computational strategies, in particular non-intrusive methods such as interpolation, and obtain a similar convergence rate ?

(iii) Can we obtain similar approximation results for more general models, including nonlinear PDEs with non-affine dependence in the parameters, and still with parametric dimension $d = \infty$?

This thesis is motivated by these questions and brings precise answers to all of them.

4 Outline of the thesis

This thesis is composed of seven chapters and comprises three main parts numbered I,II and III. Most of the results presented in the thesis are published in our papers [22, 26, 25, 21, 24, 23].

Part I deals with theoretical approximation results, while parts II and III deal with the construction of practical approximation algorithms. In Part I (chapters 1-2), we recall the results in [34] and [33] for the linear elliptic PDE (1.6) with affine parametric dependence (1.7) and present our results from [25] that apply to more general models including nonlinear PDEs with non-affine parametric dependence. In part II (chapters 3-4), we present intrusive algorithms for the approximation of the solution map u of (1.6), either by mean of adaptively constructed Taylor series, following our approach from [22], or by mean of Galerkin projections, in the line of [53]. In part III (chapters 5-6-7), we discuss two non-intrusive algorithms that can be used for the treatment of general parametric PDEs: interpolation, following our approach developed in [26, 21, 24] and least squares, following our approach from [23]. It should be noted that throughout the thesis, we mostly place ourself in the infinite dimensional setting,

$$d = \infty \qquad \text{and} \qquad U := [-1,1]^{\mathbb{N}}, \tag{4.1}$$

but the various algorithms and results are also applicable in the finite dimensional setting $d < \infty$.

In Chapter 1, we study the parametric elliptic problem (1.6) with affine dependence (1.7) and uniform ellipticity assumption (1.8). For this model, we recall in detail the approximation results obtained in [34, 33]. As previously explained, theses results show that under the assumption that $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathcal{F})$ for some $p < 1$, the Taylor series truncated to the sets Λ_n of their n largest terms converge toward the solution map u in the uniform sense with algebraic rate n^{-s} where $s := \frac{1}{p} - 1$, and the Legendre series converge in the mean square sense with the rate n^{-s^*} where $s^* := \frac{1}{p} - \frac{1}{2}$. In addition, we establish that the best n-term sets $(\Lambda_n)_{n\geq 0}$ which are used to obtain such rates can be modified so that they are lower sets in the sense of (3.8), while still maintaining the same convergence rate. This is of importance for the construction and convergence study of numerical algorithms in Part II and III. Although Chapter 1 gathers to a large extent the results obtained in [34, 33], it is self contained and can be red without referring to the cited papers. We have undertaken the task of explaining thoroughly the paradigm of treating the infinite dimension, yet shortening and simplifying the reasonings from [34, 33]. The analysis in Chapter 1 is essential in understanding the various tools that are used in further chapters.

In Chapter 2, we study parametric PDEs of the general form (1.1) with anisotropic dependences on the parameters y_j, however not necessarily of the type (1.6) with affine dependence (1.7), following our approach from [25]. As a toy example, consider the model (1.6) with diffusion coefficient

$$a(x,y) := \overline{a} + \Big(\sum_{j\geq 1} y_j \psi_j\Big)^2. \tag{4.2}$$

Although uniform ellipticity in $y \in U$ is maintained for this new form for a, the Taylor series may not anymore converge. In fact, this is already the case when all the functions ψ_j are equal to 0 except $\psi_1 = b > 1$ constant and $\overline{a}$ constant equal to 1. Indeed, in this case

$$u(y) = \frac{u(0)}{1 + b^2 y_1^2}, \qquad y \in U, \tag{4.3}$$

is a function for which the Taylor expansion diverges on $[-1, 1]$. However, this function remains the sum of its Legendre series. Following the approach of Chapter 1 for the approximation by Legendre polynomials, we show that a large class of parametric PDEs can be approximated with similar algebraic convergence rates n^{-s} by n-term truncations of Legendre series. For this purpose, we introduce the notion of (p, ε)-holomorphy which describes certain anisotropy assumptions that guarantee the algebraic convergence rate. We introduce two general framework in which (p, ε)-holomorphy holds, the first one based on Inf-Sup conditions and the second one on the holomorphic version of the implicit function theorem in Banach spaces. We show that these frameworks apply to various parametric PDEs. In particular, we consider semi-linear elliptic problems of the form

$$g(u) - \operatorname{div}(a(y)\nabla u) = f, \tag{4.4}$$

parabolic problems such as

$$\partial_t u - \operatorname{div}(a(y)\nabla u) = f, \tag{4.5}$$

and PDEs set on domains which themselves depend on the parameter y.

The remainder of the thesis, Parts II and III, is dedicated to the design of practical algorithms for constructing polynomial approximations to parametric PDEs. Given a family of polynomials $\mathcal{P}$ as in (3.3), two main problems need to be addressed:

- The identification of good index sets Λ_n in the sense u can be well approximated in $\mathbb{V}_{\Lambda_n}(\mathcal{P})$.
- The practical computation of a good approximation $u_{\Lambda_n} \in \mathbb{V}_{\Lambda_n}(\mathcal{P})$ to u.

As stressed several times, it is of the utmost importance that both tasks are numerically tractable. We will see that in both the uniform and mean average sense, it is of interest to rely on adaptive algorithms. For this type of algorithms, the sequence of index sets $(\Lambda_n)_{n\geq 1}$ is not known in advance, the identification of every set Λ_{n+1} is only based on the available information gained from computation at the previous step n. In particular, we rigorously demonstrate the effectiveness of such type of algorithms in chapters 3-4 which are concerned with the elliptic problem (1.6).

Chapter 3 discusses our results from [22] which deal with the approximation of u by computable Taylor series. In [34, 33], it is proved that Taylor coefficients, defined in (3.38), are the unique solutions of the following recursive problems

$$\int_D \overline{a}(x)\nabla t_\nu(x)\nabla w(x)dx = -\sum_{j:\nu_j\neq 0}\int_D \psi_j(x)\nabla t_{\nu-e_j}(x)\nabla w(x)dx, \qquad w\in V, \tag{4.6}$$

where $e_j = (\delta_{i,j})_{i\geq 1}$ is the Kronecker sequence of index j and $\nu - e_j$ the subtraction element-wise of e_j from ν, that is,

$$\nu - e_j = (\nu_1, \ldots, \nu_{j-1}, \nu_j - 1, \nu_{j+1}, \ldots). \tag{4.7}$$

Building upon this recursion and using residual reduction arguments of the same type as introduced in adaptive wavelet methods [30, 31], we show that one can incrementally construct a sequence of nested sets $(\Lambda_k)_{k\geq 1}$ such that the associated Taylor series converge with the optimal rate n^{-s} with $n = n(k) := \#(\Lambda_k)$ proved in Chapter 1. In view of (4.6), the Taylor series associated with a given set Λ can be computed in exactly $\#(\Lambda)$ recursion steps if and only if

$$\nu \in \Lambda \Rightarrow \nu - e_j \in \Lambda \text{ for any } j \geq 1 \text{ such that } \nu_j \neq 0. \tag{4.8}$$

This definition is equivalent to (3.8). The index sets that are adaptively built for Taylor series are therefore always lower sets.

In chapter 4, we investigate the approximation of u by Galerkin projection in the mean square sense, using the Legendre polynomials $(L_\nu)_{\nu\in\mathcal{F}}$. We only consider the case where the joint probability ϱ of y is the uniform measure in U. However, all the results can be readily extended to other product measures, if one replaces Legendre polynomials by the orthogonal tensorized polynomials for the given product measure. Following the approach of [53], we formulate the variational problem (3.14) in the Legendre basis, which yields an equivalent sequence space formulation

$$\mathbf{A}\mathbf{u} = \mathbf{f} \tag{4.9}$$

where $\mathbf{A} = (\mathbf{A}_{\nu,\nu'})_{\nu,\nu'\in\mathcal{F}}$ an infinite matrix of operators defined from V to V^*, $\mathbf{u} = (v_\nu)_{\nu\in\Lambda} \in \ell^2(\mathcal{F}, V)$ the sequence of Legendre coefficients of u and $\mathbf{f} = (f_\nu)_{\nu\in\Lambda} \in \ell^2(\mathcal{F}, V^*)$ the sequence of Legendre coefficients of $f \in (L^2(U, V, d\varrho))^*$. Using the machinery of adaptive wavelet methods for elliptic operator equations [30, 31, 49], we study the properties of the infinite matrix $\mathbf{A}$, then by a residual analysis we incrementally build a nested sequence of index sets $(\Lambda_k)_{k\geq 1}$ such that the Galerkin projection $\mathbf{u}_{\Lambda_k} \in \ell^2(\Lambda_k, V)$ of the previous formulation converge to $\mathbf{u}$ with the prescribed rate n^{-s^*} with $n = n(k) := \#(\Lambda_k)$ proved in Chapter 1. In contrast with Chapter 3, the sets Λ_k are not necessarily lower sets. We show that similar approximations results can be obtained with the lower sets constraint, and that in such case, Galerkin projection also approximate u in the uniform sense.

The methods presented in chapters 3-4 are intrusive. They are specifically designed for the elliptic linear problem (1.6) with affine dependence as in (1.7). In particular, the convergence analysis is strongly tied to these features. For more general models, these methods might be difficult to apply and one may prefer to rely on non-intrusive methods. The latter become unavoidable in cases where one has no complete knowledge on the PDE and only has access to the solution $u(y)$ for any query y through a numerical solver. In this direction, we investigate in part III two frequently used non-intrusive methods, namely interpolation and least squares.

In chapter 5, we present the interpolation process that we introduced in [26]. The process is defined through a generalisation of the Smolyak formula defining the interpolation operators (3.30) for the isotropic and anisotropic simplices $\mathcal{S}_k$ and $\mathcal{S}_{k,\alpha}$, now replaced by arbitrary lower index sets. We generalize in particular the results of [9, 7] in which it is shown that for certain types of lower sets Λ, the operator $\mathcal{I}_\Lambda$ defined by (3.30) with simply Λ replacing $\mathcal{S}_k$ is an interpolation operator. To be specific, given $(r_0, r_1, \ldots)$ a sequence of pairwise distinct points in $[-1, 1]$ and denoting by I_k the polynomial interpolation operator associated with $(r_0, \ldots, r_k)$, with the convention $I_{-1} = 0$, then for any lower set Λ,

$$\mathcal{I}_\Lambda := \sum_{i\in\Lambda} \otimes_{j=1}^d (I_{i_j} - I_{i_j-1}), \tag{4.10}$$

is the interpolation onto $\mathbb{P}_\Lambda$ for the grid of points

$$\Gamma_\Lambda := \Big\{ r_\nu := (r_{\nu_j})_{j\geq 1} : \nu \in \Lambda \Big\}. \tag{4.11}$$

We show that such operators can be computed easily by a Newton-like formula. Namely, given Λ a lower set and $\nu \in \mathcal{F} \setminus \Lambda$ such that $\Lambda' := \Lambda \cup \{\nu\}$ is lower, then

$$\mathcal{I}_{\Lambda'} u = \mathcal{I}_\Lambda u + \Big(u(r_\nu) - \mathcal{I}_\Lambda u(r_\nu)\Big) h_\nu, \tag{4.12}$$

where

$$h_\nu(y) = \prod_{j \geq 1} h_{\nu_j}(y_j), \quad \text{with} \quad h_0 = 1 \quad \text{and} \quad h_k(t) := \prod_{j=0}^{k-1} \frac{t - r_j}{r_k - r_j}. \tag{4.13}$$

We then study the stability of the interpolation through the analysis of Lebesgue constants $\mathbb{L}_\Lambda := \|\mathcal{I}_\Lambda\|_{L^\infty \to L^\infty}$. We show in particular that the growth of such constants in terms of the size of Λ can be estimated from the growth of the Lebesgue constants $\mathbb{L}_k$ associated with the operator I_k. More precisely, we prove

$$\mathbb{L}_k \leq (k+1)^\theta \quad \text{for any} \quad k \geq 1 \quad \Rightarrow \quad \mathbb{L}_\Lambda \leq (\#\Lambda)^{\theta+1} \quad \text{for any lower set } \Lambda. \tag{4.14}$$

The polynomials growth $(\#(\Lambda))^{\theta+1}$ might be larger than the algebraic decays $(\#(\Lambda))^{-s}$ that we established in chapters 1-2 for the approximation of PDEs by polynomials in the spaces $\mathbb{V}_\Lambda$. However, we show under the same assumptions as those in the results of chapters 1-2 that there exist a sequence of lower sets $(\Lambda_n)_{n \geq 0}$ with $\#(\Lambda_n) = n$, such that the approximation to u by $\mathcal{I}_{\Lambda_n} u$ converges at the optimal rate n^{-s}, with $s = \frac{1}{p} - 1$.

We also use formula (4.12) as a starting point to the development of adaptive algorithms where, for a given Λ_n, the newly chosen ν such that $\Lambda_{n+1} = \Lambda_n \cup \{\nu\}$ maximizes the increment $\Big(u(r_\nu) - \mathcal{I}_\Lambda u(r_\nu)\Big) h_\nu$ in some norm of interest. Although such adaptive algorithms are not proved to converge with the optimal rate, they appear to behave quite well in several relevant test cases. Finally, we extend the idea of sparse high dimensional interpolation to other tensorized systems than polynomials, in particular piecewise affine and quadratic functions, based on similar concepts of lower sets.

Motivated by the result expressed by (4.14), we are interested in finding infinite sequences $(r_0, r_1, \ldots)$ such that the Lebesgue constants $\mathbb{L}_k$ associated to the sections $(r_0, \ldots, r_k)$ have moderate algebraic growth. Note that Chebychev or Gauss-Lobatto points result in a logarithmic growth of the Lebesgue constant, however such points are not the sections of a single infinite sequence. In chapter 6, we study the growth of the Lebesgue constant associated with the so-called Leja sequences on the unit complex disk and their projection into $[-1, 1]$ the so-called $\Re$-Leja sequences. This chapter is a follow up of our paper [21] and two anterior works [18, 19] in which these sequences are studied. We provide new structural properties of these sequences, then prove a new bound on the growth of Lebesgue constant of $\Re$-Leja sequences. Namely

$$\mathbb{L}_k \leq 8\sqrt{2}(k+1)^2, \quad k \geq 0, \tag{4.15}$$

which improves the bound $8(k+1)^2 \log(k+1)$ that we established in [21]. This new result, shows in particular that starting with an $\Re$-Leja sequence $(r_0, r_1, \ldots)$, the resulting high dimensional interpolation operator $\mathcal{I}_\Lambda$ has Lebesgue constant bounded by $(\#(\Lambda))^3$ regardless of the dimension d and of the shape of Λ.

In chapter 7, we present the results of our paper [23] in which the stability of polynomial least squares in high dimension is investigated. Given Λ a lower set of cardinality n and $\mathcal{O}_m := (y^i, z^i)_{i=1,\ldots,m}$, where the y^i are i.i.d. copies of the random parameter vector y and z^i are noiseless or noisy observations of the solution map u at y^i, the least squares projection is defined by

$$\mathcal{I}_{\Lambda,\mathcal{O}_m} u := \operatorname*{argmin}_{v \in \mathbb{V}_\Lambda} \frac{1}{m} \sum_{i=1}^{m} \|z^j - v(y^j)\|_V^2. \tag{4.16}$$

When V is a Hilbert space, the solution of this problem is obtained by solving a simple linear systems similar to the case of real valued data. Using techniques from [32], we investigate the stability of the least squares projection, in terms of a compromise between the dimension $n = \#(\Lambda)$ of the polynomial space $\mathbb{V}_\Lambda$ and the sample size m. In particular, when ϱ is the uniform measure on U, we show that the projection is stable for values of m that scale at least like n^2.

Part I

Sparse polynomial approximation of parametric PDEs

Chapter 1

Elliptic PDEs with affine parameter dependance

Contents

1.1 Introduction

In this chapter, we study the parametric elliptic problem (1.6) associated with diffusion coefficients with affine dependence on an infinite dimensional vector y as in (1.11).

In order to facilitate referencing, we provide again the description of this parametric problem: given D a bounded Lipschitz domain of $\mathbb{R}^m$, $m \geq 1$, and $f \in H^{-1}(D)$, we consider the parametrized family of elliptic boundary value problems

$$-\text{div}(a(y)\nabla u) = f \quad \text{in} \quad D, \quad u = 0 \quad \text{on} \quad \partial D. \tag{1.1.1}$$

Here, for $y := (y_j)_{j\geq 1}$ ranging in $U := [-1,1]^{\mathbb{N}}$, the diffusion function $a(y)$ is defined over D by

$$a(y)(x) := \overline{a}(x) + \sum_{j\geq 1} y_j \psi_j(x), \quad x \in D, \tag{1.1.2}$$

where $\overline{a}$ and $(\psi_j)_{j\geq 1}$ are functions of $L^\infty(D)$, such that the above series converges absolutely for any $x \in D$ and any $y \in U$. For notational simplicity, we use the notation $a(x,y)$ instead of $a(y)(x)$. We assume that the parametric diffusion coefficient a satisfies the *uniform ellipticity assumption*

$$\mathbf{UEA}(r,R): \qquad 0 < r \leq a(x,y) \leq R < \infty, \quad y \in U, \quad x \in D. \tag{1.1.3}$$

The previous parametric problem is considered in the study of steady state diffusion problems which are subject to internal diffusivity uncertainties (in the stochastic setting) or controls (in the deterministic setting). For example, fluid diffusion in random heterogeneous porous media, heat diffusion in domain with random thermal conductivity, control of the heat flux through the design of a thermal component, etc.

By Lax-Milgram theory, $\mathbf{UEA}(r,R)$ ensures for every $y \in U$ the existence and uniqueness of the solution $u(y)$ of (1.1.1) in the space

$$V := H_0^1(D). \tag{1.1.4}$$

Moreover, the solutions $u(y)$ are uniformly bounded according to

$$\sup_{y\in U} \|u(y)\|_V \leq \frac{\|f\|_{V^*}}{r}. \tag{1.1.5}$$

This chapter is in large part discussing the approximation results obtained in the papers [34, 33] for the approximation of the solution map defined from U to V by

$$u : y \mapsto u(y), \tag{1.1.6}$$

using sparse Taylor and Legendre series. Previously to the mentioned works, the parametric elliptic model was extensively studied in the literature, e.g. [39, 3, 2, 82, 5, 6, 65, 74, 4, 70, 69, 68, 7, 8] and references their-in.

As discussed in the general introduction of this thesis, these papers propose various strategies based on multivariate polynomials in the setting $d < \infty$. Specific polynomial approximation methods are considered, for instance Galerkin projection or collocation

on pre-defined polynomial spaces, then the convergence of the built polynomial approximations is investigated. For the purposes of the latter task, as for scalar-valued functions, the smoothness of u in the parametric variable y is examined. The approach of [34, 33] is quite different, in the sense that the authors investigate the rate of polynomial approximation of u in dimension $d = \infty$ for some optimal truncations of polynomial expansions. Such optimal truncations are practically out of reach, yet they may serve as benchmark for the convergence of adaptive and non-adaptive algorithms, as discussed in part II and III of the thesis. The analysis in [34, 33] is also based on the study of the smoothness of u in the parametric variable y. However, in order to be able to treat the infinite dimensional setting $d = \infty$, it is crucial to assume a certain form of decay in the size of the ψ_j as $j \to +\infty$. Such decay is intuitively due to the fact that the convergence of the affine series (1.1.2) for all $y \in U$ should typically be reflected by a certain form of decay in the size of ψ_j as $j \to +\infty$, resulting into weaker dependence in the corresponding variables y_j. As a consequence, the discretization tools should also reflect this anisotropy. We recall in ample details a paradigm introduced in [34, 33] for dealing with this purpose and give new results in the same lines of work.

First, we show in §1.2, using arguments of best n-term approximations, how one can overcome the curse of dimensionality in polynomial approximation of the solution map u. More precisely, we explain that when u is equal to its Taylor series and the sequence $(\|t_\nu\|_V)$ of V-norm of Taylor coefficients is ℓ^p-summable with $0 < p < 1$, then truncated Taylor series associated with the n largest Taylor coefficients converge towards u with algebraic rates $(n+1)^{-s}$, where $s := \frac{1}{p} - 1$, in the uniform sense. In the mean square sense, we show that the same approach applied with Legendre series yield convergence towards u with algebraic rates $(n+1)^{-s^*}$, where $s^* := \frac{1}{p} - \frac{1}{2}$.

Motivated by the analysis of §1.2, we examine in §1.3 a first approach for the study of the summability of Taylor and Legendre coefficients of u, based on a recursive computation of the partial derivatives. Recalling the arguments of [34], we show in both cases that, for $0 < p < 1$, the ℓ^p-summability of the coefficients sequences is inherited from a similar property of ℓ^p-summability for the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$, under the additional assumption that this sequence has sufficiently small ℓ^1-norm.

We then discuss in §1.4 a more powerful approach, introduced in [33], based on extending u to the complex domain. More precisely, we show that u has an holomorphic extension to complex polydiscs with variable radii in each direction y_j reflecting the anisotropy on the dependence in y. Based on this anisotropic holomorphy, we derive new estimates on the V-norm of Taylor coefficients by application of Cauchy formula. This allow us to prove that they are ℓ^p-summable under the only assumption that $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$,

In §1.5, we introduce the notion of lower set which was already given in the general introduction, see (3.8), and study its interplay with best n term approximation. We show in particular that the estimates obtained for Taylor coefficients by complex

analysis in §1.4 are actually better than ℓ^p-summable in the sense that they allow us to replace the best n-term index sets by lower sets of same cardinality n, while maintaining the rate $(n+1)^{-s}$.

Finally, in §1.6, we investigate the approximation of the solution map u by its Legendre series or more generally by its *Jacobi series*. Using again complex analysis arguments as in [33], we first recall the result stating that, for $0 < p < 1$, Legendre coefficients are ℓ^p-summabe under the only condition that $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathbb{N})$. Then, we prove a new theorem stating that the same result holds in general with Jacobi coefficients and that, similarly to Taylor coefficients, the summability is stronger than ℓ^p in the sense that the best n-term index sets can be replaced by lower sets, while maintaining the convergence rate $(n+1)^{-s^*}$.

1.2 Sparse best n-term polynomial approximation

A natural obstruction when approximating the solution map u by multivariate polynomials in y is the difficulty that y is infinite dimensional. In particular, the use of standard polynomial spaces, such as of degree at most k in each variable, is not appropriate since such spaces would then be infinite dimensional. This also reflects the fact that these spaces are not well adapted to finite yet high dimensional problems, since their dimension grows exponentially with the number d of variables.

The approach introduced in [34, 33] consists instead in building *sparse* polynomial approximations. These approximations are obtained by keeping the largest terms in polynomial expansions of the map u. These expansions are either the Taylor series of u at $y = 0$ or the Legendre series. Other expansions are discussed in §1.6.

We introduce $\mathcal{F}$ the set of *finitely supported* nonnegative integers, that consists of all $\nu := (\nu_j)_{j\geq 0}$ such that $\nu_j \in \mathbb{N}$ and $\#\{j : \nu_j \neq 0\} < \infty$. We introduce the notations

$$0_{\mathcal{F}} := (0, 0, \ldots), \tag{1.2.1}$$

for the null sequence, and

$$\nu! := \prod_{j\geq 1} \nu_j! \quad \text{and} \quad y^\nu := \prod_{j\geq 1} y_j^{\nu_j}, \qquad \nu \in \mathcal{F}, \quad y \in U, \tag{1.2.2}$$

with $0! = 0^0 = 1$. The Taylor expansion of u is formally defined as the series

$$u(y) = \sum_{\nu\in\mathcal{F}} t_\nu y^\nu, \qquad t_\nu := \frac{\partial^\nu u(0)}{\nu!} \in V \tag{1.2.3}$$

The existence of the derivatives $\partial^\nu u(0)$ and the rate of convergence of this series towards u are discussed further in §1.3 and §1.4. For now, we assume that the solution map u is equal to its Taylor series for any $y \in U$.

We next introduce the Legendre expansion of u, using two possible normalizations. We denote by $(P_n)_{n\geq 0}$ and $(L_n)_{n\geq 0}$ the Legendre polynomials over $[-1,1]$ normalised in the uniform and the mean square sense respectively, i.e.

$$\|P_n\|_{L^\infty([-1,1])} = P_n(1) = 1, \qquad \|L_n\|^2_{L^2([-1,1],\frac{dt}{2})} = \int_{-1}^{1} L_n(t)^2 \frac{dt}{2} = 1. \tag{1.2.4}$$

We recall that the family $(L_n)_{n\geq 0}$ is an orthonormal basis in the space $L^2([-1,1],\frac{dt}{2})$. Since $\|L_n\|_{L^\infty[-1,1]} = \sqrt{2n+1}$, the polynomials P_n and L_n are related according to

$$P_n = \frac{L_n}{\sqrt{2n+1}}, \quad n \geq 0. \tag{1.2.5}$$

We now define the multivariate polynomials $(L_\nu)_{\nu\in\mathcal{F}}$ and $(P_\nu)_{\nu\in\mathcal{F}}$ by tensorization

$$L_\nu(y) := \prod_{j\geq 1} L_{\nu_j}(y_j) \quad \text{and} \quad P_\nu(y) := \prod_{j\geq 1} P_{\nu_j}(y_j) \qquad \nu \in \mathcal{F}, \quad y \in U. \tag{1.2.6}$$

The products are well defined and finite since $L_0 = P_0 = 1$. We denote by ϱ the uniform measure over U, i.e

$$d\varrho(y) := \otimes_{j\geq 1} \frac{dy_j}{2}. \tag{1.2.7}$$

The sigma algebra Θ for $d\varrho$ is generated by the finite rectangles $\prod_{j=1}^{\infty} S_j$ where only a finite number of the S_j are different for $[-1,1]$ and are intervals contained in $[-1,1]$. Here $(U,\Theta,d\varrho)$ is a probability space. It is easy to check that the family $(L_\nu)_{\nu\in\mathcal{F}}$ forms an orthonormal system in $L^2(U,d\varrho)$. Moreover, this system is complete. Indeed, any function of $L^2(U,d\varrho)$ can be approximated to any given tolerance by a finite linear combination of characteristic function of finite rectangle and each of the last can be approximated by polynomials to any prescribed accuracy.

The family of polynomials $(L_\nu)_{\nu\in\mathcal{F}}$ constitute an orthonormal basis for the Bochner space

$$\mathcal{V}_2 := L^2(U,V,d\varrho), \tag{1.2.8}$$

of square ϱ-measurable mappings from U to V equipped with the least-square norm

$$\|u\|_{\mathcal{V}_2} := \Big(\int_U \|u(y)\|_V^2 d\varrho(y)\Big)^{\frac{1}{2}}. \tag{1.2.9}$$

We also introduce the space

$$\mathcal{V}_\infty \subset L^\infty(U,V), \tag{1.2.10}$$

of functions u defined *everywhere* in U and uniformly bounded in V, equipped with the uniform norm

$$\|u\|_{\mathcal{V}_\infty} := \sup_{y\in U} \|u(y)\|_V. \tag{1.2.11}$$

The space $\mathcal{V}_\infty$ is continuously embedded in $\mathcal{V}_2$. In view of the inequality (1.1.5), the solution map u of (1.1.1) belongs to $\mathcal{V}_\infty \subset \mathcal{V}_2$, and is therefore equal to its Legendre series

$$u(y) = \sum_{\nu\in\mathcal{F}} v_\nu L_\nu = \sum_{\nu\in\mathcal{F}} u_\nu P_\nu, \qquad v_\nu := \int_U u(y)L_\nu(y)d\varrho(y) \in V, \qquad u_\nu := \Big(\prod_{j\geq 1}(2\nu_j+1)\Big)^{\frac{1}{2}} v_\nu. \tag{1.2.12}$$

The key idea for the truncation of the above Taylor and Legendre expansions (1.2.3) and (1.2.12) comes from *nonlinear approximation* [40]. It consists in retaining the n *largest terms* in these expansions. This is a typical strategy in data compression: for example, one retains the largest wavelet coefficients of a digital image in order to encode it in an economical way. Therefore, the set $\Lambda_n \subset \mathcal{F}$ of the indices corresponding to the n retained terms is not a-priori fixed, but rather adaptively chosen with respect to the solution map.

One crucial observation is that the convergence rate of such nonlinear approximations is related to the summability of the terms in the expansion. This is expressed by the following result, originally due to Stechkin. For convenience, we formulate this lemma for sequences indexed by $\mathcal{F}$.

Lemma 1.2.1

Let $p > 0$ and $(e_\nu)_{\nu\in\mathcal{F}}$ a sequence in $\ell^p(\mathcal{F})$. If Λ_n is any set corresponding to the n largest $|e_\nu|$ then for any $q > p$

$$\Big(\sum_{\nu\notin\Lambda_n} |e_\nu|^q\Big)^{1/q} \leq \|(e_\nu)\|_{\ell^p(\mathcal{F})}(n+1)^{-s_{p,q}}, \qquad s_{p,q} = \frac{1}{p} - \frac{1}{q}. \tag{1.2.13}$$

Proof: We introduce $(e_j^*)_{j\geq 1}$ a decreasing rearrangement of the sequence $(|e_\nu|)_{\nu\in\mathcal{F}}$. On the one hand, we have

$$\Big(\sum_{\nu\notin\Lambda_n} |e_\nu|^q\Big)^{1/q} = \Big(\sum_{j\geq n+1} (e_j^*)^q\Big)^{1/q} \leq \Big(\sum_{j\geq n+1} (e_{n+1}^*)^{q-p}(e_j^*)^p\Big)^{1/q} \leq (e_{n+1}^*)^{1-p/q}\|(e_\nu)\|_{\ell^p(\mathcal{F})}^{p/q}.$$

On the other hand, we have

$$(n+1)(e_{n+1}^*)^p \leq \sum_{j=1}^{n+1}(e_j^*)^p \leq \|(e_\nu)\|_{\ell^p(\mathcal{F})}^p.$$

The combination of the two inequalities and $q \geq p$ implies (1.2.13). ∎

We note that the sets Λ_n in the previous lemma are generally not unique because of possible ties in the values of the $|e_\nu|$. However, the decay of the tail is the same for all

the index sets as it was shown in the proof. We observe also that the index sets Λ_n in the lemma may be chosen nested, that is $\Lambda_n \subset \Lambda_{n+1}$ for all n.

In light of the previous lemma, we can investigate the decay of the tail of Taylor and Legendre expansions. If we assume for example that the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ for some $0 < p < 1$, then with $(\Lambda_n^T)_{n\geq 1}$ any sequence of nested sets of indices corresponding each to the n largest $\|t_\nu\|_V$, we have

$$\Big\|u - \sum_{\nu\in\Lambda_n^T} t_\nu y^\nu\Big\|_{\mathcal{V}_\infty} \leq \sum_{\nu\notin\Lambda_n^T} \|t_\nu\|_V \leq \|(\|t_\nu\|_V)\|_{\ell^p(\mathcal{F})}(n+1)^{-s}, \qquad s = \frac{1}{p} - 1, \quad (1.2.14)$$

where in the first inequality we have used the fact that $|y^\nu| \leq 1$ for $y \in U$, and in the second inequality we have applied (1.2.13) with $q = 1$. Quite remarkably the rate $(n+1)^{-s}$ and the constant in (1.2.14) are independent from d the number of parameters y_j since we have assumed here that d is countably infinite. Thus, (1.2.14) shows that one can in principle overcome the curse of dimensionality in the approximation of u.

Similarly, if we assume that the sequences $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$, for some $0 < p < 1$ and $(\Lambda_n^L)_{n\geq 1}$ is any sequence of nested set of indices corresponding each to the n largest $\|v_\nu\|_V$, then

$$\Big\|u - \sum_{\nu\in\Lambda_n^L} v_\nu L_\nu\Big\|_{\mathcal{V}_2} = \Big(\sum_{\nu\notin\Lambda_n^L} \|v_\nu\|_V^2\Big)^{\frac{1}{2}} \leq \|(\|v_\nu\|_V)\|_{\ell^p(\mathcal{F})}(n+1)^{-s^*}, \qquad s^* = \frac{1}{p} - \frac{1}{2}, \quad (1.2.15)$$

where we have applied (1.2.13) with $q = 2$. Here again, the rate $(n+1)^{-s^*}$ and the constant are independent of the number of parameters y_j. Let us remark that, in view of the Parseval equality in (1.2.15), the series $\sum_{\nu\in\Lambda_n^L} v_\nu L_\nu$ is the best possible n-term approximation of u by multivariate Legendre series in $\mathcal{V}_2$.

Legendre polynomials can also provide approximations in the uniform sense, that is in $\mathcal{V}_\infty$. For instance, if the sequence $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ for some $0 < p < 1$ and $(\Lambda_n^P)_{n\geq 1}$ is any sequence of nested sets of indices corresponding each to the n largest values of $\|u_\nu\|_V$, then

$$\Big\|u - \sum_{\nu\in\Lambda_n^P} u_\nu P_\nu\Big\|_{\mathcal{V}_\infty} \leq \sum_{\nu\notin\Lambda_n^P} \|u_\nu\|_V \leq \|(\|u_\nu\|_V)\|_{\ell^p(\mathcal{F})}(n+1)^{-s}, \qquad s = \frac{1}{p} - 1. \quad (1.2.16)$$

Since $\|v_\nu\|_V \leq \|u_\nu\|_V$ for any $\nu \in \mathcal{F}$, then the ℓ^p summability of $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ implies also the ℓ^p summability of $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$. However, it is important to note that the index sets Λ_n^P and Λ_n^L differ. In particular the truncated series $\sum_{\nu\in\Lambda_n^P} u_\nu P_\nu$ might not yield the optimal rate in $\mathcal{V}_2$ achieved by $\sum_{\nu\in\Lambda_n^L} v_\nu L_\nu$. Likewise, the truncated series $\sum_{\nu\in\Lambda_n^L} v_\nu L_\nu$ might not yield the optimal rate in $\mathcal{V}_\infty$ achieved by $\sum_{\nu\in\Lambda_n^P} u_\nu P_\nu$. Therefore, the truncation strategy is strongly tied to the norm in which we want to measure the error.

We are thus interested in the summability properties of the sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$, $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$. As for real valued functions, this analysis requires the study of the smoothness of the approximated function, that is the solution map $y \mapsto u(y)$, in order to derive upper bounds on the Taylor and Legendre coefficients. In the next section §1.3, we present the regularity results from [34] which are obtained by a real variable technique based on recursive differentiation of the solution map u with respect to the variables y_j. These results lead to upper bounds and summability results for the Taylor and Legendre coefficients that are suboptimal, in particular in the cases where the support of the functions ψ_j do not overlap much. This drawback is circumvented by a different approach from [33] based on holomorphy results obtained by complex variable arguments, which we present in §1.4.

1.3 Regularity and summability by the real variable technique

The function $u(y)$ is the unique solution of the variational problem

$$\int_D a(y)\nabla u(y)\nabla w = \int_D fw, \qquad w \in V, \tag{1.3.1}$$

where ∇ is the gradient with respect to the variable $x \in D$. The assumption $\mathbf{UEA}(r, R)$ implies that the solution map $y \mapsto u(y)$ is uniformly bounded in V, according to the classical a-priori estimate

$$\sup_{y\in U} \|u(y)\|_V \leq \frac{\|f\|_{V^*}}{r}. \tag{1.3.2}$$

Our next objective is to define computable approximations to the solution map $y \mapsto u(y)$ by means of polynomial expansions in the variable y with coefficients in V.

1.3.1 Differentiability of the solution map

Given $0 < r < R < \infty$, we denote $\mathcal{S}_{r,R}$ the open subset of functions in $L^\infty(D)$ that satisfies the uniform ellipticity assumption $\mathbf{UEA}(r, R)$ in (1.1.3) with strict inequalities. By Lax-Milgram theory, for every $a \in \mathcal{S}_{r,R}$, there exist a unique solution $v(a) \in V$ of the variational problem

$$\int_D a\nabla v(a)\nabla w = \int_D fw, \qquad w \in V. \tag{1.3.3}$$

In addition, the defined map $v : \mathcal{S}_{r,R} \mapsto V$ is uniformly bounded with

$$\sup_{a\in\mathcal{S}_{r,R}} \|v(a)\|_V \leq \frac{\|f\|_{V^*}}{r}. \tag{1.3.4}$$

We begin by observing that this map is Frechet differentiable.

Lemma 1.3.1

For any function a in $\mathcal{S}_{r,R}$, the map $v : a \mapsto v(a)$ if Frechet differentiable at a and its Frechet derivative $d_a v$ is defined as follows: for any $h \in L^\infty(D)$ the function $d_a v(h) \in V$ is the unique solution $\tilde{v} = \tilde{v}(a,h)$ of the variational problem

$$\int_D a\nabla\tilde{v}\nabla w = -\int_D h\nabla v(a)\nabla w, \qquad w \in V \tag{1.3.5}$$

Proof: Let a in $\mathcal{S}_{r,R}$ and $h \in L^\infty(D)$ such that $a+h \in \mathcal{S}_{r,R}$, the functions $v(a), v(a+h) \in V$ are well defined and are the unique solutions of the variational problems

$$\int_D a\nabla v(a)\nabla w = \int_D fw, \qquad \int_D (a+h)\nabla v(a+h)\nabla w = \int_D fw, \qquad w \in V.$$

Therefore for any $w \in V$

$$\int_D a\nabla(v(a+h)-v(a))\nabla w = -\int_D h\nabla v(a+h)\nabla w.$$

In particular with $w = v(a+h)-v(a)$, using **UEA**(r,R) in the left side, and (1.3.4) and Cauchy Schwartz in the right side, we deduce that

$$\|v(a+h)-v(a)\|_V \leq \frac{\|f\|_{V^*}}{r^2}\|h\|_{L^\infty}, \tag{1.3.6}$$

which shows in particular that v is a Lipchitz map. Now subtracting the variational problem (1.3.5) to the previous variational problem, we obtain

$$\int_D a\nabla(v(a+h)-v(a)-\tilde{v})\nabla w = -\int_D h\nabla(v(a+h)-v(a))\nabla w.$$

Substituting here w by $v(a+h)-v(a)-\tilde{v}$ and using **UEA**(r,R) in the left side and Cauchy-Schwartz inequality and the previous Lipchitz inequality in the right side, we obtain

$$\|v(a+h)-v(a)-\tilde{v}\|_V \leq \frac{\|f\|_{V^*}}{r^3}\|h\|^2_{L^\infty},$$

which confirms that v is Frechet differentiable at a with $d_a v(h) = \tilde{v}(a,h)$. ■

We return to the regularity of the solution map $y \mapsto u(y)$ of the parametric problem (1.1.1). We may view this map as the composition

$$u = v \circ a, \tag{1.3.7}$$

where a acts from U to the open set $\mathcal{S}_{\frac{r}{2},2R}$ of $L^\infty(D)$ according to

$$a(y) = \overline{a} + \sum_{j\geq 1} y_j \psi_j, \tag{1.3.8}$$

and where v is the previously introduced map, now defined over $\mathcal{S}_{\frac{r}{2},2R}$. By the chain rule, we have $\partial_{e_j} u(y) = d_{a(y)} v(\psi_j)$, hence $\partial_{e_j} u(y) \in V$ is the unique solution of the variational problem

$$\int_D a(y) \nabla \partial_{e_j} u(y) \nabla w = -\int_D \psi_j \nabla u(y) \nabla w, \qquad w \in V. \tag{1.3.9}$$

Note that this variational problem can be directly obtained by formal differentiation of (1.3.1) with respect to y. This reasoning can be repeated in order to prove the existence of higher order derivatives $\partial_\nu u(y)$ for any $\nu \in \mathcal{F}$. The variational problem satisfied by the derivative $\partial_\nu u(y)$ is obtained by deriving "ν times" with respect to y the variational problem (1.3.1). Using Leibniz formula in multi-dimension and the affine dependence of a in the parameter y, we find that for $\nu \neq 0_{\mathcal{F}}$ the derivative $\partial_\nu u(y) \in V$ is the unique solution of the variational problem

$$\int_D a(y) \nabla \partial_\nu u(y) \nabla w = -\sum_{j:\nu_j \neq 0} \nu_j \int_D \psi_j \nabla \partial_{\nu - e_j} u(y) \nabla w, \qquad w \in V. \tag{1.3.10}$$

Let us note that the sum in the right hand side is finite for any $\nu \in \mathcal{F} \setminus \{0\}$. The previous formula implies that Taylor coefficients t_ν satisfy recursive formulas. Indeed, setting $y = 0$ and dividing by $\nu!$, one obtains the variational problem

$$\int_D \overline{a} \nabla t_\nu \nabla w = -\sum_{j:\nu_j \neq 0} \int_D \psi_j \nabla t_{\nu - e_j} \nabla w, \qquad w \in V. \tag{1.3.11}$$

The explicit recursive formulas (1.3.10) and (1.3.11) allows us to obtain a priori estimates on the V-norms $\|\partial_\nu u(y)\|_V$ which shall provide us with a preliminary understanding of the summability of the the Taylor and Legendre coefficients.

1.3.2 Upper estimates of Taylor and Legendre coefficients

The previous recursive formulas can be used in order to bound the partial derivatives $\partial_\nu u(y)$, and subsequently the Taylor and Legendre coefficients. We introduce the sequence

$$b := (b_j)_{j\geq 1}, \quad b_j := \|\psi_j\|_{L^\infty(D)}. \tag{1.3.12}$$

Considering the formula (1.3.10) with $w = \partial_\nu u(y)$ and applying the uniform ellipticity assumption **UEA**(r, R) and Cauchy Schwartz inequality, we obtain that for any $y \in U$

$$\|\partial_\nu u(y)\|_V \leq \sum_{j:\nu_j \neq 0} \nu_j \frac{b_j}{r} \|\partial_{\nu - e_j} u(y)\|_V, \qquad \nu \neq 0_{\mathcal{F}}. \tag{1.3.13}$$

Since by (1.3.2), the map $\partial_{0_{\mathcal{F}}} u = u$ is uniformely bounded by $\frac{\|f\|_{V^*}}{r}$, then an immediate induction yields

$$\|\partial_\nu u\|_{\mathcal{V}_\infty} \leq \frac{\|f\|_{V^*}}{r} |\nu|! \; c^\nu, \qquad \nu \in \mathcal{F}, \tag{1.3.14}$$

where we have used the notation

$$|\nu| := \sum_{j:\nu_j \neq 0} \nu_j, \tag{1.3.15}$$

and introduced the sequence $c = (c_j)_{j\geq 1} := (\frac{b_j}{r})_{j\geq 1}$. The same justification, based on the recursive formulas (1.3.11), shows that Taylor coefficients satisfy the following

$$\|t_\nu\|_V \leq \frac{\|f\|_{V^*}}{\bar{r}} \frac{|\nu|!}{\nu!} \; \bar{c}^\nu, \tag{1.3.16}$$

where $\bar{c} = (\bar{c}_j)_{j\geq 1} := (\frac{b_j}{\bar{r}})_{j\geq 1}$ with $\bar{r} := \min_{x\in D} \bar{a}(x)$. As for the Legendre coefficients, using Rodrigues formulas $P_n(t) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dt^n}\{(1-t^2)^n\}$, we obtain from (1.3.10) by inductive integration by parts in the variables y_j that the Legendre coefficients defined in (1.2.12) satisfy

$$v_\nu := \frac{1}{\nu!} \prod_{j:\nu_j\neq 0} \frac{\sqrt{2\nu_j+1}}{2^{\nu_j}} \int_U \partial_\nu u(y) \prod_{j:\nu_j\neq 0} (1-y_j^2)^{\nu_j} d\varrho(y) \tag{1.3.17}$$

Therefore, these coefficients can be bounded according to

$$\|v_\nu\|_V \leq \frac{\|\partial_\nu u\|_{\mathcal{V}_\infty}}{\nu!} \prod_{j:\nu_j\neq 0} I_{\nu_j}, \qquad I_n := \frac{\sqrt{2n+1}}{2^n} \int_{-1}^{1} (1-t^2)^n \; \frac{dt}{2}. \tag{1.3.18}$$

The sequence $(I_n)_{n\geq 1}$ can be computed explicitly and shown to satisfy $I_n \leq (\frac{1}{\sqrt{3}})^n$, the inequality being sharp since for $n = 1$ we have $I_1 = \frac{1}{\sqrt{3}}$. We deduce then that

$$\|v_\nu\|_V \leq \frac{\|f\|_{V^*}}{r} \frac{|\nu|!}{\nu!} \; \tilde{c}^\nu, \qquad \|u_\nu\|_V = \beta_\nu \|v_\nu\|_V \leq \frac{\|f\|_{V^*}}{r} \beta_\nu \frac{|\nu|!}{\nu!} \; \tilde{c}^\nu, \tag{1.3.19}$$

where we have defined $\tilde{c} = (\tilde{c}_j)_{j\geq 1} := (\frac{b_j}{r\sqrt{3}})_{j\geq 1}$ and the sequence $(\beta_\nu)_{\nu\in\mathcal{F}}$ with $\beta_\nu = \prod_{j\geq 1} \sqrt{2\nu_j+1}$.

The bounds (1.3.16) and (1.3.19) have been obtained by cumulating a series of inequalities, and are therefore expected not to be sharp. As explained further in §1.4, these bounds are particularly overestimated when the supports of the functions ψ_j do not overlap, and in this case better bounds can be obtained by a complex variable technique.

We may however produce an example for which the bound (1.3.16) is sharp. In this example, we assume that the functions $\overline{a}$ and ψ_j are constants. We assume for example that $\overline{a}$ take the value 1 and denote by $-b_j < 0$ the value of each function ψ_j. The uniform ellipticity assumption $\mathbf{UEA}(r,R)$ is then equivalent to $b := (b_j)_{j\geq 1} \in \ell^1(\mathbb{N})$ with $r \leq 1 - \|b\|_{\ell^1(\mathbb{N})} \leq 1 + \|b\|_{\ell^1(\mathbb{N})} \leq R$ or simply $\|b\|_{\ell^1(\mathbb{N})} \leq 1 - r$ and the value $2 - r$ for R. Since the diffusion coefficient is contant over D, the solution map u is simply given by

$$u(y) = \frac{u_0}{1 - \sum_{j=1}^{\infty} b_j y_j}, \qquad y \in U. \tag{1.3.20}$$

Here, $u_0 := u(0) \in V$ is the unique solution of the Laplace equation

$$-\Delta v = f \quad \text{in} \quad D, \quad v = 0 \quad \text{on} \quad \partial D. \tag{1.3.21}$$

Since we may write

$$(1 - \sum_{j=1}^{\infty} b_j y_j)^{-1} = \sum_{k\geq 0} \Big(\sum_{j=1}^{\infty} b_j y_j \Big)^k = \sum_{k\geq 0} \sum_{|\nu|=k} \frac{k!}{\nu!} b^\nu y^\nu = \sum_{\nu\in\mathcal{F}} \frac{|\nu|!}{\nu!} b^\nu y^\nu, \tag{1.3.22}$$

then the Taylor coefficients are given by $t_\nu = \frac{|\nu|!}{\nu!} b^\nu u_0$. This shows that in this particular case, the bound (1.3.16) is sharp up to a multiplicative constant.

1.3.3 Summability of upper estimates

The summability properties of the sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ can be investigated using the upper bounds in (1.3.16) and (1.3.19). In both cases, the analysis amounts to the study of a sequence of the form $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ with α a sequence of positive real number. We observe that $\alpha = (\alpha_j)_{j\geq 1}$ is a subsequence of $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ associated with the indices e_j, hence if α is not $\ell^p(\mathbb{N})$-summable then $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ is not $\ell^p(\mathcal{F})$-summable. It is then of interest to investigate what condition one should assume on α beside the ℓ^p-summability that implies the ℓ^p-summability of $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ with the same $p > 0$. This was done in [34] for $0 < p < 1$. We give the result and its proof in Theorem 1.3.2 below. We begin by giving the intermediate result [34, Lemma 7.1] formulated with $0 < p \leq 1$ in [34] but which we remark is valid also for any $p > 0$, as shown in the proof.

Theorem 1.3.2

Let $\alpha := (\alpha_j)_{j\geq 1} \in [0, +\infty[^{\mathbb{N}}$. For any $p > 0$, the sequence $(\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $\alpha \in \ell^p(\mathbb{N})$ and $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$. Moreover

$$\|(\alpha^\nu)\|_{\ell^p(\mathcal{F})} \leq \exp\Big(\kappa_p \frac{\|\alpha\|^p_{\ell^p(\mathbb{N})}}{p}\Big), \qquad \kappa_p := \frac{1}{1 - \|\alpha\|^p_{\ell^\infty(\mathbb{N})}}. \tag{1.3.23}$$

Proof: It is sufficient to establish the result for $p = 1$, since one can consider $\alpha^p = (\alpha_j^p)_{j\geq 1}$ in the case $p \neq 1$. We assume that $(\alpha^\nu)_{\nu\in\mathcal{F}} \in \ell^1(\mathcal{F})$, then α which can be seen as $(\alpha^{e_j})_{j\geq 1}$ belongs to $\ell^1(\mathbb{N})$. Moreover, from the identity

$$\sum_{\nu\in\mathcal{F}} \alpha^\nu = \prod_{j\geq 1}\sum_{n\geq 0} \alpha_j^n, \tag{1.3.24}$$

it is necessary that $\alpha_j < 1$ for every $j \geq 1$. This combined with $\lim_{j\to\infty} \alpha_j = 0$ implies that necessarily $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$ and settles the "only if" implication. For the "if" implication, we have by (1.3.24)

$$\sum_{\nu\in\mathcal{F}} \alpha^\nu = \prod_{j\geq 1}\Big(1 + \frac{\alpha_j}{1-\alpha_j}\Big) \leq \prod_{j\geq 1} \exp\Big(\frac{\alpha_j}{1-\alpha_j}\Big) \leq \prod_{j\geq 1} \exp(\kappa_1\alpha_j), \tag{1.3.25}$$

where we have applied the inequality $1 + t \leq e^t$, valid for any $t \in \mathbb{R}$, with the real numbers $\frac{\alpha_j}{1-\alpha_j}$. We thus find that the sequence $(\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^1(\mathcal{F})$ with its ℓ^1-norm dominated by $\exp(\kappa_1\|\alpha\|_{\ell^1(\mathbb{N})})$. The bound (1.3.23) for $p \neq 1$ is obtained by considering $(\alpha_j^p)_{j\geq 1}$ which belongs to $\ell^1(\mathbb{N})$. ■

The condition $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$ in the previous theorem yields a decay of the sequence $(\alpha^\nu)_{\nu\in\mathcal{F}}$ which allows us to obtain the summability result. It is of interest to show that this summability is not affected by the presence of certain type of algebraic factors, as expressed by the following theorem, which is not given in [34]. We introduce a notation for polynomial growth in multi-dimension by: for $C > 0$ and $\theta \geq 0$ real numbers, we define the sequence $(C_\nu(\theta))_{\nu\in\mathcal{F}}$ by

$$C_{0_\mathcal{F}}(\theta) = 1 \quad \text{and} \quad C_\nu(\theta) := \prod_{j:\nu_j\neq 0} C\,\nu_j^\theta \quad \text{for} \quad \nu \in \mathcal{F} - \{0\}. \tag{1.3.26}$$

Theorem 1.3.3

Let $\alpha := (\alpha_j)_{j\geq 1} \in [0, +\infty[^{\mathbb{N}}$, $C > 0$ and $\theta \geq 0$. For any $p > 0$, the sequence $\Big(C_\nu(\theta)\alpha^\nu\Big)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $\alpha \in \ell^p(\mathbb{N})$ and $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$. In addition

$$\|(C_\nu(\theta)\alpha^\nu)\|_{\ell^p(\mathcal{F})} \leq \exp\Big(m!C^p\kappa_p^{m+1}\frac{\|\alpha\|_{\ell^p(\mathbb{N})}^p}{p}\Big), \tag{1.3.27}$$

with κ_p as in Theorem 1.3.2 and $m = \theta p$ if $\theta p \in \mathbb{N}$ or $m = \lceil \theta p \rceil$ otherwise.

Proof: Up to work with α^p, C^p and $p\theta$, it suffices to prove the theorem for $p = 1$. If the sequence $(C_\nu(\theta)\alpha^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^1(\mathcal{F})$, then the subsequence $C\alpha = (C_{e_j}(\theta)\alpha^{e_j})_{j\geq 1}$ is in $\ell^1(\mathbb{N})$ and so is α. In addition, by the identity

$$\sum_{\nu\in\mathcal{F}} C_\nu(\theta)\alpha^\nu = \prod_{j\geq 1}\Big(1 + \sum_{n\geq 1} Cn^\theta\alpha_j^n\Big),$$

necessarily $\alpha_j < 1$ for any $j \geq 1$. This combined with $\alpha_j \to 0$ completes the "only if" implication. In order to prove the "if" implication, we let $m = \theta$ if $\theta \in \mathbb{N}$ and $m = \lceil \theta \rceil$ otherwise, so that $n^\theta \leq n^m \leq \frac{(n-1+m)!}{(n-1)!}$ for any $n \geq 1$. We have then for $t \in]0,1[$

$$\sum_{n\geq 1} n^\theta t^n \leq \sum_{n\geq 1} \frac{(n-1+m)!}{(n-1)!} t^n = t\Big(\sum_{n\geq 0} t^{n+m}\Big)^{(m)} = t\Big(\frac{t^m}{1-t}\Big)^{(m)} = \frac{m!\, t}{(1-t)^m}.$$

Therefore

$$\sum_{\nu\in\mathcal{F}} C_\nu(\theta)\alpha^\nu \leq \prod_{j\geq 1}\Big(1+C\frac{m!\alpha_j}{(1-\alpha_j)^{m+1}}\Big) \leq \prod_{j\geq 1}\exp\Big(C\frac{m!\alpha_j}{(1-\alpha_j)^{m+1}}\Big) \leq \prod_{j\geq 1}\exp(m!C\kappa_1^{m+1}\alpha_j),$$

which completes the proof. The bound (1.3.3) is obtained for $p \neq 0$ using α^p, C^p and $p\theta$. ∎

We now turn to the ℓ^p summability of $(\frac{|\nu|!}{\nu!}\alpha^\nu)_{\nu\in\mathcal{F}}$ which, in view of $\frac{|\nu|!}{\nu!} \geq 1$ for any ν, will necessarily demands a stronger condition than $\|\alpha\|_{\ell^\infty(\mathbb{N})} < 1$. The following result is given in [34].

Theorem 1.3.4

Let $\alpha := (\alpha_j)_{j\geq 1} \in [0,+\infty[^{\mathbb{N}}$. For any $0 < p < 1$, the sequence $\Big(\frac{|\nu|!}{\nu!}\alpha^\nu\Big)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $\alpha \in \ell^p(\mathbb{N})$ and $\sum_{j\geq 1}\alpha_j < 1$.

Proof: We assume that $\Big(\frac{|\nu|!}{\nu!}\alpha^\nu\Big)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F}) \subset \ell^1(\mathcal{F})$, then α which is the subsequence associated with the indices e_j is in $\ell^p(\mathbb{N})$. Moreover, the quantity

$$\sum_{\nu\in\mathcal{F}} \frac{|\nu|!}{\nu!}\alpha^\nu = \sum_{k=0}^{\infty}\Big(\sum_{j\geq 1}\alpha_j\Big)^k, \tag{1.3.28}$$

is finite. Therefore, necessarily $\sum_{j\geq 1}\alpha_j < 1$ and we have proved the "only if" implication. The "if" implication is not straightforward, due to the fact that no identity similar to (1.3.28) is available for $\sum_{\nu\in\mathcal{F}}(\frac{|\nu|!}{\nu!}\alpha^\nu)^p$ with $p \neq 1$. We decompose the sequence α into the product $\alpha_j = \gamma_j\delta_j$ with γ and δ to be precised later. By Holder inequality, we have

$$\sum_{\nu\in\mathcal{F}}\Big(\frac{|\nu|!}{\nu!}\alpha^\nu\Big)^p = \sum_{\nu\in\mathcal{F}}\Big(\frac{|\nu|!}{\nu!}\gamma^\nu\Big)^p(\delta^\nu)^p \leq \Big\|\Big(\frac{|\nu|!}{\nu!}\gamma^\nu\Big)\Big\|^p_{\ell^1(\mathcal{F})}\|(\delta^\nu)\|^p_{\ell^{p'}(\mathcal{F})},$$

where $p' = \frac{p}{1-p}$. In view of Theorem 1.3.2 and the first part of the proof, the right side in the above inequality is finite whenever $\gamma \in \ell^1(\mathbb{N})$ with $\|\gamma\|_{\ell^1(\mathbb{N})} < 1$ and $\delta \in \ell^{p'}(\mathbb{N})$ with $\|\delta\|_{\ell^\infty(\mathbb{N})} < 1$. We now construct a factorization of α into two sequences γ and δ that satisfy these properties. For $\eta > 0$ and $J \geq 1$ to be specified later, we define δ by

$$\delta_j = \frac{1}{1+\eta}, \quad \text{for} \quad 1 \leq j \leq J-1, \qquad \delta_j = \alpha_j^{1-p}, \quad \text{for} \quad j \geq J.$$

We have $\delta \in \ell^{p'}(\mathbb{N})$ and $\|\delta\|_{\ell^\infty(\mathbb{N})} < 1$ because $\alpha \in \ell^p(\mathbb{N})$ and $0 \le \alpha_j < 1$ for any $j \ge 1$. The sequence γ is given by

$$\gamma_j = (1+\eta)\alpha_j, \quad \text{for} \quad 1 \le j \le J-1, \qquad \delta_j = \alpha_j^p, \quad \text{for} \quad j \ge J.$$

The sequence γ is in $\ell^1(\mathbb{N})$ because $\alpha \in \ell^p(\mathbb{N})$. Moreover, we have

$$\|\gamma\|_{\ell^1(\mathbb{N})} = (1+\eta)\sum_{j\ge 1}^{J-1} \alpha_j + \sum_{j \ge J+1} \alpha_j^p.$$

Since $\|\alpha\|_{\ell^1(\mathbb{N})} < 1$, then taking η such that $1+\eta < \frac{1}{\|\alpha\|_{\ell^1(\mathbb{N})}}$ and J large enough, we get $\|\gamma\|_{\ell^1(\mathbb{N})} < 1$, which concludes the proof. ■

Similar to Theorem 1.3.3, the summability is not affected by the presence of algebraic factors of the type $C_\nu(\theta)$.

Theorem 1.3.5

Let $\alpha := (\alpha_j)_{j\ge 1} \in [0,+\infty[^{\mathbb{N}}$, $C \ge 1$ and $\theta \ge 0$. For $0 < p < 1$, the sequence $\left(C_\nu(\theta)\frac{|\nu|!}{\nu!}\alpha^\nu\right)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ if and only if $\alpha \in \ell^p(\mathbb{N})$ and $\sum_{j\ge 1}\alpha_j < 1$.

Proof: It is similar to the proof of the previous theorem. The "only if" implication follows from the observation $\frac{|\nu|!}{\nu!}\alpha^\nu \le C_\nu(\theta)\frac{|\nu|!}{\nu!}\alpha^\nu$ because we assumed $C \ge 1$. For the "if" implication, we use the decomposition $C_\nu(\theta)\frac{|\nu|!}{\nu!}\alpha^\nu = (\frac{|\nu|!}{\nu!}\gamma^\nu)(C_\nu(\theta)\delta^\nu)$, then use the same choice for γ and δ and apply Theorem 1.3.3. ■

In light of the previous Theorems 1.3.4 and 1.3.5, we are now able to study the ℓ^p-summability of the estimates in (1.3.16) and (1.3.19). If those estimates belong to $\ell^p(\mathcal{F})$ for some $0 < p < 1$, then necessarily the sequences $\bar{c}$ and $\tilde{c}$ belong to $\ell^p(\mathbb{N})$ with the same p. This obviously holds if and only if $(\|\psi_j\|_{L^\infty})_{j\ge 1}$ belongs to $\ell^p(\mathbb{N})$. In view of the multiplicative constants relating the sequences $\bar{c}$, $\tilde{c}$ and b, we thus obtain the following result.

Theorem 1.3.6

Under the uniform ellipticity assumption and if the sequence $b = (\|\psi_j\|_{L^\infty})_{j\ge 1}$ belongs to $\ell^p(\mathbb{N})$ for some $0 < p < 1$, then

- *If $\|b\|_{\ell^1(\mathbb{N})} < \bar{r} = \min_{x\in D}\bar{a}(x)$, the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$.*
- *If $\|b\|_{\ell^1(\mathbb{N})} < \sqrt{3}r$, the sequences $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$.*

For the sequences $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$, we have used directly the estimates (1.3.16) and (1.3.19) and Theorem 1.3.4. Concerning the sequence $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$, we have

used Theorem 1.3.5, taking into account the algebraic coefficients β_ν in (1.3.19). This last result was not given in [34].

Note that we have the implication

$$\|b\|_{\ell^1(\mathbb{N})} < \bar{r} \Rightarrow \mathbf{UEA}(r, R) \quad \text{with} \quad r := \bar{r} - \|b\|_{\ell^1(\mathbb{N})},\ R = \|\overline{a}\|_{L^\infty(R)} + \|b\|_{\ell^1(\mathbb{N})}. \tag{1.3.29}$$

In the particular case where the functions $\overline{a}$ and ψ_j are constants, $\|b\|_{\ell^1} < \bar{r} = \overline{a}$ is exactly equivalent to $\mathbf{UEA}(r, R)$ with such values of r and R. However, in the more general case of non-constant $\overline{a}$ and ψ_j, the condition $\|b\|_{\ell^1} < \bar{r}$ could be much stronger than $\mathbf{UEA}(r, R)$. This is particularly clear when the supports of ψ_j do not overlap too much, since in that case the maximal value of $\sum_{j\geq 1} |\psi_j(x)|$ for $x \in D$ may be much smaller than $\|b\|_{\ell^1}$. In that sense, restrictions on the value of $\|b\|_{\ell^1}$ such as in the above theorem appear artifical. In the next section, we discuss a different approach for the estimation of Taylor and Legendre coefficients, which leads to ℓ^p summability results, under the sole assumptions that $b = (\|\psi_j\|_{L^\infty})_{j\geq 1}$ belongs to $\ell^p(\mathbb{N})$ and that $\mathbf{UEA}(r, R)$ holds, without such artificial restrictions.

1.4 Holomorphy of the solution map on the complex variable

1.4.1 Holomorphy of the solution map

We describe shortly the process of extending the solution map u of the parametric problem (1.1.1) to certain regions of $\mathbb{C}^{\mathbb{N}}$. The steps of this process are more detailed in [33]. First, we extend the parametrization of the diffusion coefficient a according to

$$a(z) := \overline{a} + \sum_{j\geq 1} z_j \psi_j, \quad z \in \mathbb{C}^{\mathbb{N}}. \tag{1.4.1}$$

We have that for every $z \in C^{\mathbb{N}}$ such that $\sup_{j\geq 1} |z_j| < \infty$, the coefficient $a(z)$ is well defined and belongs to $L^\infty(D)$ the space of complex valued and bounded functions over D. Indeed, the uniform ellipticity assumption $\mathbf{UEA}(r, R)$ is equivalent to

$$0 < r \leq \overline{a}(x) - \sum_{j\geq 1} |\psi_j(x)| \leq \overline{a}(x) + \sum_{j\geq 1} |\psi_j(x)| \leq R < \infty, \qquad x \in D, \tag{1.4.2}$$

therefore, for z as above, we have $\|a(z)\|_{L^\infty(D)} \leq R \max\{1, \sup_{j\geq 1} |z_j|)\}$. We introduce the notation

$$\mathcal{U} := \otimes_{j\geq 1}\{|z_j| \leq 1\}, \tag{1.4.3}$$

for the unit poly-disc in $\mathbb{C}^{\mathbb{N}}$. In view of the previous discussion, the diffusion coefficient a is well defined on $\mathcal{U}$. Moreover, using the uniform ellipticity assumption $\mathbf{UEA}(r, R)$

recalled in the form (1.4.2), we have

$$\textbf{UEAC}(r,R): \qquad 0<r\leq \Re(a(x,z))\leq |a(x,z)|\leq R<\infty, \quad z\in\mathcal{U}, \quad x\in D. \tag{1.4.4}$$

By the complex version of Lax-Milgram theory, for any $z\in\mathcal{U}$, there exists a unique solution $u(z)$ of the elliptic problem (1.1.1) with parameter z. This function $u(z)$ belongs to V the complex version of the sobolev space $H_0^1(D)$ and is the unique solution of the variational problem

$$\int_D a(z)\nabla u(z)\overline{\nabla w}=\int_D f\overline{w}, \qquad w\in V. \tag{1.4.5}$$

Moreover, as with the real variable, the uniform ellipticity inequality $\textbf{UEAC}(r,R)$ in (1.4.4) implies that the solution map $u: z\in\mathcal{U}\mapsto u(z)$ is uniformly bounded in V with

$$\sup_{z\in\mathcal{U}}\|u(z)\|_V\leq \frac{\|f\|_{V^*}}{r}. \tag{1.4.6}$$

The map $u: z\in\mathcal{U}\mapsto u(z)$ is an extension of the map $u: y\in U\mapsto u(y)$. It is fundamental to check that this extension is holomorphic with respect to every variable z_j. In other words, u is continuously $\mathbb{C}$-differentiable with respect to every variable z_j. The linearity of the parametric variational problem (1.4.5) and the affine dependence of the complex diffusion coefficient a in z suggests using the same perturbation arguments of the previous section 1.3 used for the real variable.

Given $0<r<R<\infty$, we denote this time by $\mathcal{S}_{r,R}$ the open subset of functions in $L^\infty(D)$ that satisfies the uniform ellipticity assumption $\textbf{UEAC}(r,R)$ in (1.4.4) with strict inequalities. By Lax-Milgram theory, for every $a\in\mathcal{S}_{r,R}$, there exists a unique solution $v(a)\in V$ of the variational problem

$$\int_D a\nabla v(a)\overline{\nabla w}=\int_D f\overline{w}, \qquad w\in V. \tag{1.4.7}$$

In addition, the defined map $v:\mathcal{S}_{r,R}\mapsto V$ is uniformly bounded with

$$\sup_{a\in\mathcal{S}_{r,R}}\|v(a)\|_V\leq \frac{\|f\|_{V^*}}{r}. \tag{1.4.8}$$

As for the real variable case, the map v is Frechet $\mathbb{C}$-diffentiable. We have

Lemma 1.4.1

For any function a in $\mathcal{S}_{r,R}$, the map v if Frechet $\mathbb{C}$-differentiable at a and its Frechet derivative $d_a v$ is defined as follows: for any $h\in L^\infty(D)$ the function $d_a v(h)\in V$ is

the unique solution $\tilde{v} = \tilde{v}(a, h)$ of the variational problem

$$\int_D a\nabla\tilde{v}\overline{\nabla w} = -\int_D h\nabla v(a)\overline{\nabla w}, \qquad w \in V \tag{1.4.9}$$

The proof of this lemma is exactly similar to that of Lemma 1.4.1. We can now view $u : \mathcal{U} \mapsto V$ as the composition $u = v \circ a$ where a acts from $\mathcal{U}$ to the open set $\mathcal{S}_{\frac{r}{2},2R}$ as in (1.4.1) and where v is the previously introduced map, now defined over $\mathcal{S}_{\frac{r}{2},2R}$. In particular, the holomorhy of a on z combined with the previous lemma implies that u admits partial derivatives of first order $\partial_{e_j}u(z)$ at any $z \in \mathcal{U}$. By the chain rule, we have $\partial_{e_j}u(z) = d_{a(z)}v(\psi_j)$, hence $\partial_{e_j}u(z) \in V$ is the unique solution of the variational problem

$$\int_D a(z)\nabla\partial_{e_j}u(z)\overline{\nabla w} = -\int_D \psi_j\nabla u(z)\overline{\nabla w}, \qquad w \in V. \tag{1.4.10}$$

The holomorphy of the solution map is then established.

In order to derive sharp estimates on Taylor and Legendre coefficients, the authors in [33] used the extension of the map u by holomorphy to neighbourhood of poly-discs wider than $\mathcal{U}$. We discuss this approach for Taylor coefficients and postpone the same approach for Legendre coefficients to Section 1.6 where the more general case of Jacobi polynomials is presented.

Definition 1.4.2

Given $0 < \delta < r$, we say that a sequence $\rho := (\rho_j)_{j\geq 1}$ is δ-admissible if and only if $\rho_j \geq 1$ for any $j \geq 1$ and

$$\sum_{j\geq 1}(\rho_j - 1)|\psi_j(x)| \leq r - \delta, \qquad x \in D. \tag{1.4.11}$$

We note that there exist δ-admissible sequences ρ that satisfie $\rho_j > 1$ for every $j \geq 0$. Indeed, in view of (1.4.2), the sequence ρ with $\rho_j := 1 + \frac{r-\delta}{2^j}\frac{1}{R}$ for any $j \geq 1$ is δ-admissible. We observe that if ρ is δ-admissible then the previous inequality combined with (1.4.2) implies

$$0 < \delta \leq \overline{a}(x) - \sum_{j\geq 1}\rho_j|\psi_j(x)| \leq \overline{a}(x) + \sum_{j\geq 1}\rho_j|\psi_j(x)| \leq R + r - \delta < \infty, \qquad x \in D. \tag{1.4.12}$$

This is equivalent to say that a satisfies the uniform ellipticity assumption $\mathbf{UEAC}(\delta, R+r-\delta)$ over the domain

$$\mathcal{U}_\rho := \otimes_{j\geq 1}\{|z_j| \leq \rho_j\}. \tag{1.4.13}$$

Using exactly the same arguments as above, it can be proved that the map u can be extended by holomorphy to the domain $\mathcal{U}_\rho$ with for any $z \in \mathcal{U}_\rho$, $u(z)$ is the unique solution of (1.4.5). Similarly to (1.4.6), the map u stays uniformly bounded in the domain $\mathcal{U}_\rho$ with

$$\sup_{z\in\mathcal{U}_\rho} \|u(z)\|_V \leq \frac{\|f\|_{V^*}}{\delta}. \tag{1.4.14}$$

The above discussion shows that, under only the uniform ellipticity assumption (1.4.2), it is possible to extend the map u by holomorphy to wider polydiscs $\mathcal{U}_\rho$ than the unit polydisc $\mathcal{U}$, with the possibility that $\mathcal{U}_{\rho_j}$ contains strictly $\{|z_j| \leq 1\}$ for any $j \geq 1$. The same method can be carried again, in order to prove that u can be extended beyond the domain $\mathcal{U}_\rho$. For instance, given $\delta \in]0, r[$ and ρ a δ-admissible sequence, it is easily checked that the sequence ρ', defined by $\rho'_j := \rho_j + \frac{\delta/2}{2^j}\frac{1}{R}$, is $\{\frac{\delta}{2}\}$-admissible, so that the map u can be extended by holomorphy to the domain $\mathcal{U}_{\rho'}$. We note that individually, the interior of every disk $\mathcal{U}_{\rho'_j}$ is a neighbourhood of the disk $\mathcal{U}_{\rho_j}$.

Remark 1.4.3

The holormorphy properties of the map $z \mapsto u(z)$ allows us to justify the fact that the Taylor series $\sum_{\nu\in\mathcal{F}} t_\nu z^\nu$ converges uniformly towards u over the polydisc $\mathcal{U}$, under the assumption that $b \in \ell^1(\mathbb{N})$. Indeed, on the hand, for any $J \geq 0$, the map

$$(z_1, \ldots, z_J) \mapsto u(z_1, \ldots, z_J, 0, 0, \ldots), \tag{1.4.15}$$

is holomorphic in a neighbourhood of the finite dimensional polydisc $\otimes_{j=1}^J \{|z_j| \leq 1\}$, which implies the uniform convergence of its Taylor series. On the other hand, using the property (1.3.6), we find that

$$\sup_{z\in\mathcal{U}} \|u(z) - u(z_1, \ldots, z_J, 0, 0, \ldots)\|_V \leq \frac{\|f\|_{V^*}}{r^2} \sum_{j>J} b_j, \tag{1.4.16}$$

which tends to 0 as $J \to \infty$. This shows that the Taylor series converges according to certain summability processes which progressively activate the variables z_j as j grows. However, the results on ℓ^p summability of the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ for $p \leq 1$ imply that the series converges unconditionally, see [33] for more details on this point.

1.4.2 Upper bounds and summability of the Taylor coefficients

We let δ be in $]0, r[$, ρ a δ-admissible sequence and ρ' a $\{\frac{\delta}{2}\}$-admissible such that the interior of every $\mathcal{U}_{\rho'_j}$ is a neighbourhood of $\mathcal{U}_{\rho_j}$. We consider $\nu \neq 0$ a multi-index in $\mathcal{F}$ and assume without loss of generality that $\{1, \ldots, J\}$ is the support of ν, that is $\nu_j \neq 0$

for $j = 1, \dots, J$ and $\nu_j = 0$ for $j \geq J+1$. The Taylor coefficient t_ν defined in (1.2.3) is equal to the Taylor coefficient associated with the index $(\nu_1, \dots, \nu_J) \in \mathbb{N}^J$ of the map u_J defined over $[-1,1]^J$ by $u_J : (y_1, \dots, y_J) \mapsto u(y_1, \dots, y_J, 0, 0, \dots)$. This last map can be extended by holomorphy to the poly-disk $\otimes_{j=1}^J \mathcal{U}_{\rho'_j}$. Since the interior of every $\mathcal{U}_{\rho'_j}$ is a neighbourhood of $\mathcal{U}_{\rho_j}$, then by successive application of Cauchy integral formula on the variables $z_1, \dots, z_J$, we obtain

$$t_\nu = (2i\pi)^{-J} \int_{|z_1|=\rho_1} \dots \int_{|z_J|=\rho_J} \frac{u(z_1, \dots, z_J, 0, 0, \dots)}{z_1^{\nu_1} \dots z_J^{\nu_J}} dz_1 \dots dz_J. \tag{1.4.17}$$

Using the uniform bound (1.4.14), the previous bound implies $\|t_\nu\|_V \leq \frac{\|f\|_{V^*}}{\delta} \prod_{j=1}^J \rho_j^{-\nu_j}$. We introduce the notation $\rho^{-\nu} := \prod_{j:\nu_j \neq 0} \rho_j^{-\nu_j}$ for $\nu \in \mathcal{F} - \{0\}$ and $\rho^{-0_\mathcal{F}} = 1$. The above discussion shows that for any $\delta \in]0, r[$ and any δ-admissible sequence ρ, one has

$$\|t_\nu\|_V \leq \frac{\|f\|_{V^*}}{\delta} \rho^{-\nu}, \qquad \nu \in \mathcal{F}, \tag{1.4.18}$$

The bound is valid for $0_\mathcal{F}$ because $\|t_{0_\mathcal{F}}\|_V = \|u(0)\|_V \leq \frac{\|f\|_{V^*}}{r}$. We thus have established the following result.

Theorem 1.4.4

Under the uniform ellipticity assumption **UEA**(r, R), *for any* $\nu \in \mathcal{F}$

$$\|t_\nu\|_V \leq h_\nu := \inf_{0<\delta<r} \left\{ \frac{\|f\|_{V^*}}{\delta} \inf\{\rho^{-\nu} : \rho \text{ is } \delta\text{-admissible}\} \right\} \tag{1.4.19}$$

It is not obvious if the previous estimates on Taylor coefficients are better than the estimates in (1.3.16). In order to draw a rough comparison, let us define the sequence $c := (c_j)_{j\geq 1}$ with $c_j = \frac{2\|\psi_j\|_{L^\infty(D)}}{r}$. Given an index $\nu \neq 0$, it is easily checked that the sequence $\rho = \rho(\nu)$ defined by $\rho_j = 1 + \frac{1}{c_j} \frac{\nu_j}{|\nu|}$ is $\{r/2\}$-admissible, therefore

$$\|t_\nu\|_V \leq \frac{\|f\|_{V^*}}{r/2} \rho^{-\nu} \leq 2 \frac{\|f\|_{V^*}}{r} c^\nu \frac{|\nu|^{|\nu|}}{\nu^\nu} \leq 2 \frac{\|f\|_{V^*}}{r} (ec)^\nu \frac{|\nu|!}{\nu!} \tag{1.4.20}$$

where we have used the Stirling type inequalities $n! \leq n^n \leq e^n n!$. We have then retrieved from (1.4.19) an estimate of the type obtained by the real variable techniques in (1.3.16). We remark that this estimate does not sharpen the estimate in (1.3.16), the quantity $1/\overline{r}$ being replaced with the larger quantity $2e/r$.

However, the flexibility in the choice of the sequence ρ allows us to prove better summability result on the sequence $(h_\nu)_{\nu\in\mathcal{F}}$. For instance, if the function ψ_j have mutually disjoint supports then ρ is δ-admissible if and only if $(\rho_j - 1)b_j \leq r - \delta$ for

any $j \geq 0$. Therefore, for any $\nu \neq 0$, we find in this case that

$$h_\nu = \inf_{0<\delta<r} \Big\{ \frac{\|f\|_{V^*}}{\delta} \prod_{j\geq 0} \Big(1 + \frac{r-\delta}{b_j}\Big)^{-\nu_j} \Big\} \leq \frac{\|f\|_{V^*}}{r/2} c^\nu, \quad c := (\frac{b_j}{b_j + \frac{r}{2}})_{j\geq 1}. \tag{1.4.21}$$

If the sequence b belongs to $\ell^p(\mathbb{N})$, then c also belongs to $\ell^p(\mathbb{N})$. Moreover $\|c\|_{\ell^\infty(\mathbb{N})} < 1$, then according to Theorem 1.3.2, the sequence $(c^\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$. Therefore, in this case the sequences $(h_\nu)_{\nu\in\mathcal{F}}$ and $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belong to $\ell^p(\mathcal{F})$.

One main result established in [33] shows that a similar result holds in the general case.

Theorem 1.4.5

Under the uniform ellipticity assumption $\mathbf{UEA}(r, R)$ *and if the sequence* $b = (\|\psi_j\|_{L^\infty})_{j\geq 1}$ *belongs to* $\ell^p(\mathbb{N})$ *for some* $0 < p < 1$, *then the sequences* $(h_\nu)_{\nu\in\mathcal{F}}$ *and* $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ *belong to* $\ell^p(\mathcal{F})$.

Proof: Let $J \geq 1$ be an integer to be fixed later. We write $\mathbb{N} = E \cup F$ with $E := \{1 \leq j \leq J\}$ and $F := \{j \geq J+1\}$ and introduce for $\nu = (\nu_j)_{j\geq 1} \in \mathcal{F}$ the notations $\nu_E := (\nu_1, \ldots, \nu_J) \in \mathbb{N}^J$ and $\nu_F := (\nu_{J+1}, \nu_{J+2}, \ldots) \in \mathcal{F}$. Let $\nu \in \mathcal{F} - \{0\}$ fixed. We introduce the sequence $\rho(\nu) := (\rho_j)_{j\geq 1}$ that depends on ν according to

$$\rho_j := \kappa \quad \text{for} \quad j \in E \quad \text{and} \quad \rho_j := \kappa + \frac{r/4}{|b_j|} \frac{\nu_j}{|\nu_F| + 1} \quad \text{for} \quad j \in F,$$

where $\kappa = 1 + \frac{r/4}{\|(b_j)\|_{\ell^1}}$. We have

$$\sum_{j\geq 1} (\rho_j - 1)|b_j| \leq (\kappa - 1) \sum_{j\geq 1} b_j + \frac{r}{4} \sum_{j>J} \frac{\nu_j}{|\nu_F| + 1} = \frac{r}{4} + \frac{r}{4} \frac{|\nu_F|}{|\nu_F| + 1} \leq \frac{r}{2} = r - \frac{r}{2}.$$

The sequence $\rho(\nu)$ is thus $\{r/2\}$-admissible, hence $h_\nu \leq 2\frac{\|f\|_{V^*}}{r} \rho^{-\nu}$. It is then sufficient to prove that the sequence $(q_\nu := \rho(\nu)^{-\nu})_{\nu\in\mathcal{F}}$ with $q_{0_\mathcal{F}} = 1$ belongs to $\ell^p(\mathcal{F})$. We have

$$q_\nu = q_E(\nu) q_F(\nu), \qquad q_E(\nu) := \prod_{j\leq J : \nu_j \neq 0} \kappa^{-\nu_j}, \qquad q_F(\nu) := \prod_{j>J : \nu_j \neq 0} \rho_j^{-\nu_j}.$$

We use the convention $q_E(0) = q_F(0) = 1$. We denote $\mathcal{F}_E$ the multi-indices in $\mathcal{F}$ supported in E and $\mathcal{F}_F$ the multi-indices in $\mathcal{F}$ supported in F, with convention that $0_\mathcal{F}$ belongs to both sets. The separable form of the q_ν above allows us to write

$$\sum_{\nu\in\mathcal{F}} q_\nu^p = A_E A_F \quad \text{where} \quad A_E := \sum_{\nu\in\mathcal{F}_E} q_E(\nu)^p \quad \text{and} \quad A_F := \sum_{\nu\in\mathcal{F}_F} q_F(\nu)^p.$$

On the one hand, in view of $\kappa > 1$, we have

$$A_E = \sum_{\nu\in\mathbb{N}^J} \prod_{1\leq j\leq J} \kappa^{-p\nu_j} = \Big(\sum_{n=0}^{\infty} \kappa^{-pn}\Big)^J < +\infty,$$

On the other hand, introducing the sequence $d := (d_j)_{j\geq 1}$ defined by $d_j := \frac{b_{j+J}}{r/4}$ and for $\nu \in \mathcal{F}_F$ denoting $\mu := \nu_F = (\nu_{J+1}, \nu_{J+2}, \ldots) \in \mathcal{F}$, we may write

$$q_F(\nu) \leq \prod_{j\geq 1:\mu_j\neq 0} \Big(\frac{1+|\mu|}{\mu_j} d_j\Big)^{\mu_j} = \frac{(1+|\mu|)^{|\mu|}}{\mu^\mu} d^\mu.$$

Using the Stirling type inequalities $n! \leq n^n \leq (1+n)^n \leq n! e^{n+1}$ valid for any $n \geq 1$, it follows that $q_F(\nu) \leq e \frac{|\mu|!}{\mu!}(ed)^\mu$, which is also valid for $\mu = \nu_F = 0_{\mathcal{F}}$, hence

$$A_F \leq e^p \sum_{\mu\in\mathcal{F}} \Big(\frac{|\mu|!}{\mu!}(ed)^\mu\Big)^p$$

Since $b \in \ell^1$, choosing J large enough, we may assume that

$$\|ed\|_{\ell^1(\mathbb{N})} = \frac{4e}{r} \sum_{j>J} |b_j| < 1.$$

By Theorem 1.3.4 we have $A_F < \infty$, which concludes the proof. ∎

The estimates $(h_\nu)_{\nu\in\mathcal{F}}$ that we obtained using complex analysis allows us to obtain the ℓ^p summability of the Taylor coefficients under the only condition that $\mathbf{UEA}(r, R)$ holds and that the sequence $b = (\|\psi_j\|_{L^\infty})_{j\geq 1}$ belongs to $\ell^p(\mathbb{N})$ with $0 < p < 1$, with no restrictions on the ℓ^1 norm of b. The approximation of the map u by best n-term Taylor series discussed in section 1.2 then holds. One way to construct good n-term approximations consists in considering the sequence $(h_\nu)_{\nu\in\mathcal{F}}$ instead of $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$. For instance, if $(\Lambda_n^e)_{n\geq 1}$ is any sequence of nested set of indices corresponding each to the n largest h_ν, then

$$\Big\|u - \sum_{\nu\in\Lambda_n^e} t_\nu y^\nu\Big\|_{\mathcal{V}_\infty} \leq \sum_{\nu\notin\Lambda_n^e} \|t_\nu\|_V \leq \sum_{\nu\notin\Lambda_n^e} h_\nu \leq \|(h_\nu)\|_{\ell^p(\mathcal{F})}(n+1)^{-s}, \qquad s = \frac{1}{p} - 1. \tag{1.4.22}$$

This yields approximation to u with the same rate $(n+1)^{-s}$ as with best n-term Taylor series in (1.2.14), yet with the constant $\|(h_\nu)\|_{\ell^p(\mathcal{F})}$ which is in view of (1.4.19) larger than $\|(\|t_\nu\|_V)\|_{\ell^p(\mathcal{F})}$. One advantage in considering best n-term approximations associated with $(h_\nu)_{\nu\in\mathcal{F}}$ is that, from the definition (1.4.19), this sequence $(h_\nu)_{\nu\in\mathcal{F}}$ is monotone decreasing:

$$\mu \leq \nu \quad \Longrightarrow \quad h_\nu \leq h_\mu, \tag{1.4.23}$$

where the order relation $\leq$ is defined over indices by $\nu \leq \mu$ if and only if $\nu_j \leq \mu_j$ for any $j \geq 1$. This is easily checked since the admissible sequences ρ considered in (1.4.19) are picked in $[1, +\infty[^{\mathbb{N}}$. In view of this observation, in addition to their nestedness, we may impose constraints on the shape of the index sets Λ_n^e of the following form: if $\nu \in \Lambda_n^e$, then all the indices $\mu \leq \nu$ belong to Λ_n^e. Sets with such properties are called *"lower set"*.

1.5 Lower sets

Lower sets will play a central role throughout all the chapters of this manuscript. The following section is an introduction to this type of index set. We give some of the properties needed for this chapter. Many other interesting properties of lower sets will be progressively given in subsequent chapters.

1.5.1 Definitions and properties

We define on $\mathcal{F}$ the partial order $\leq$ by $\nu \leq \mu$ if and only if $\nu_j \leq \mu_j$ for any $j \geq 1$. The corresponding strict order $<$ is then defined by $\nu < \mu$ if and only if $\nu \leq \mu$ and $\nu_j < \mu_j$ for at least one value of j.

Definition 1.5.1

A nonempty set $\Lambda \subset \mathcal{F}$ is called lower set *if and only if*

$$\nu \in \Lambda \text{ and } \mu \leq \nu \Rightarrow \mu \in \Lambda, \tag{1.5.1}$$

or equivalently, if $\nu \in \Lambda$ then $\nu - e_j \in \Lambda$ for any $j \geq 1$ such that $\nu_j \neq 0$.

Lower sets were introduced in a variety of contexts, mainly for the interesting properties of the corresponding polynomial spaces $\mathbb{P}_\Lambda := \text{span}\{y \mapsto y^\nu : \nu \in \Lambda\}$. For instance, in the context of polynomial interpolation, such spaces were introduced in the book [58] in the special case of dimension $d = 2$ and were referred to by "polynômes pleins". They were also introduced in the theory of "least polynomial space" for interpolation of functions on general multivariate point sets, see in particular [36]. The corresponding spaces $\mathbb{P}_\Lambda$ were referred to as "order closed polynomials". Among other appellations are: down set, decreasing set, initial segment and downward closed.

Considering polynomial spaces $\mathbb{P}_\Lambda$ associated to lower sets is very natural. In particular, this allows to replace the monomials y^ν in the definition of such spaces by any other tensorized basis $Q_\nu(y) = \prod_{j\geq 1} Q_{\nu_j}(y_j)$ where $Q_0 \equiv \mathbf{1}$ and Q_k has degree exactly equal to k for every $k \geq 1$ (for examples Legendre polynomials).

In the present context of parametric PDEs, lower sets were introduced in [22] in connection with computation of quasi-optimal best n-term approximation of the map u by Taylor series. They were referred to as "monotone sets". We will explain in details this connection in this chapter and Chapter 3. We first start by giving some useful properties of lower sets.

It is obvious that the smallest lower set in $\mathcal{F}$ is $\{0_\mathcal{F}\}$ and that any nonempty lower set contains the null index $0_\mathcal{F}$. Intersections and unions of lower sets are also lower sets. Particular examples of lower sets are the "rectangular blocks" $\mathcal{B}_\nu$ defined by

$$\mathcal{B}_\nu := \{\mu \in \mathcal{F} : \mu \leq \nu\}, \qquad \nu \in \mathcal{F}. \tag{1.5.2}$$

Such lower sets are finite and have cardinality $\#(\mathcal{B}_\nu) := \prod_{j\geq 1}(\nu_j + 1)$. We say that an index ν is *maximal* in a lower set Λ if and only if there is no $\mu \in \Lambda$ satisfying $\nu < \mu$. We observe that any finite lower set has at least one maximal element. Indeed, it is easily checked that any index $\nu \in \Lambda$ with the largest ℓ^1-norm $|\nu| = \sum_{i\geq 1} \nu_i$ is maximal in Λ. The only maximal element of a block $\mathcal{B}_\nu$ is ν. In general, any finite lower set Λ in $\mathcal{F}$ is completely determined by its maximal elements according to

$$\Lambda = \bigcup_{\substack{\nu\in\Lambda \\ \nu \text{ maximal}}} \mathcal{B}_\nu \, . \tag{1.5.3}$$

We observe that given Λ a lower set and $\nu \in \Lambda$, we have that $\Lambda \setminus \{\nu\}$ is a lower set if and only if ν is maximal in Λ. This remark turn out to be useful in certain algorithmic procedures.

We now turn to the connection between lower sets and sequences indexed in $\mathcal{F}$. First, suppose we have a sequence $(h_\nu)_{\nu\in\mathcal{F}}$ of positive real number that is strictly monotone decreasing, i.e. $h_\nu < h_\mu$ for any $\nu, \mu \in \mathcal{F}$ such that $\mu < \nu$. The sets $(\Lambda_n)_{n\geq 1}$ of indices corresponding each to the n largest values of h_ν are unique, lower sets and nested. Indeed, there exists a unique decreasing rearrangement $(h_{\nu^k})_{j\geq 1}$ of $(h_\nu)_{\nu\in\mathcal{F}}$, hence the sets $\Lambda_n = \{\nu^1, \dots, \nu^n\}$ are unique and nested. Also for $k \geq 2$ and $\mu < \nu^k$, we have $h_{\nu^k} < h_\mu$, so that necessarily $\mu = \nu^j$ for some $j = 1, \dots, k-1$, proving the sets are lower sets.

The previous result is not true when $(h_\nu)_{\nu\in\mathcal{F}}$ is only monotone decreasing. Many realizations of $(\Lambda_n)_{n\geq 1}$ may exist and may not be necessarily lower sets. However, there exists at least one realization $(\Lambda_n)_{n\geq 1}$ consisting in nested lower sets. In view of the structure of this realization, the index ν such that $\Lambda_{n+1} = \Lambda_n \cup \{\nu\}$ satisfies $\nu - e_j \in \Lambda_n$ for any $j \geq 1$ with $\nu_j \neq 0$. This hints how the desired realisation can then be obtained.

Algorithm 1.5.2

- *Set* $\Lambda_1 := \{\mu^1 := 0_\mathcal{F}\}$. *For* $k \geq 1$ *do;*
- Λ_k *has been defined, compute* $\mathcal{N}(\Lambda_k) = \{\nu \notin \Lambda_k : \nu - e_j \in \Lambda_k \text{ for any } j \text{ s.t } \nu_j \neq 0\}$;
- *Get* $\mu^k = \operatorname{argmax}_{\mu\in\mathcal{N}(\Lambda_k)}(e_\mu)$ *and set* $\Lambda_{k+1} = \Lambda_k \cup \{\mu^k\}$;

The sets of adjacent neighbours $\mathcal{N}(\Lambda_k)$ are of infinite cardinalities, however the argmax problems always have a solution. Indeed, the sets $\{h_\nu : \nu \in \mathcal{N}(\Lambda_k)\}$ are countable and bounded from above by $h_{0_\mathcal{F}}$. We need to verify that every Λ_k is lower set and corresponds to the largest k value of the sequence $(h_\nu)_{\nu\in\mathcal{F}}$. The first claim is true by construction. Now, given that a set Λ_k has been constructed, we let $\mu \notin \Lambda_k$. It is easily checked that $\mathcal{N}(\Lambda_k) \cap \mathcal{B}_\mu = \mathcal{N}(\Lambda_k \cap \mathcal{B}_\mu)$, hence the intersection is not empty

and we can pick $\nu \in \mathcal{N}(\Lambda_k)$ such that $\nu \leq \mu$. From the definition of μ^k, we have $h_\mu \leq h_\nu \leq h_{\mu^k}$. This shows that $(h_{\mu^k})_{k\geq 1}$ is a decreasing arrangement of $(h_\nu)_{\nu\in\mathcal{F}}$ and affirms the claim.

In the rest of this manuscript, we refer to the following definition.

Definition 1.5.3

Let $(h_\nu)_{\nu\in\mathcal{F}}$ be a monotone decreasing sequence of positive numbers. We call a lower realization associated with $(h_\nu)_{\nu\in\mathcal{F}}$ any sequence $(\Lambda_n)_{n\geq 1}$ of nested lower sets corresponding each to the n largest values of h_ν.

The notion of lower realization is particularly useful in the best n-term approximation by lower sets. We explain here this type of approximation. Given a sequence $c := (c_\nu)_{\nu\in\mathcal{F}}$, we call the monotone envelope of the sequence c, the sequence $\mathbf{c} := (\mathbf{c}_\nu)_{\nu\in\mathcal{F}}$ defined by

$$\mathbf{c}_\nu = \sup_{\nu\leq\mu} |c_\mu|. \tag{1.5.4}$$

The sequence $\mathbf{c}$ is the smallest monotone decreasing sequence bounding element-wise the sequence $(|c_\nu|)_{\nu\in\mathcal{F}}$. For $p > 0$, we introduce the space $\ell^p_m(\mathcal{F})$ of sequence indexed in $\mathcal{F}$ and defined by

$$(c_\nu)_{\nu\in\mathcal{F}} \in \ell^p_m(\mathcal{F}) \quad \text{if and only if} \quad (\mathbf{c}_\nu)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F}), \tag{1.5.5}$$

equipped with the quasi-norm

$$\|c\|_{\ell^p_m(\mathcal{F})} := \|\mathbf{c}\|_{\ell^p(\mathcal{F})}. \tag{1.5.6}$$

We note that if the sequence $c = (c_\nu)_{\nu\in\mathcal{F}}$ is monotone decreasing, then it coincides with its monotone envelope, in which case it suffices to have $c \in \ell^p(\mathcal{F})$ in order to assert that $c \in \ell^p_m(\mathcal{F})$. We note also that if c and c' are two sequence such that $|c_\nu| \leq |c'_\nu|$ for any $\nu \in \mathcal{F}$, then $c' \in \ell^p_m(\mathcal{F})$ implies $c \in \ell^p_m(\mathcal{F})$.

The following lemma is the counterpart of Lemma 1.2.1 for sequences in the spaces $\ell^p_m(\mathcal{F})$.

Lemma 1.5.4

Let $p > 0$ and $c := (c_\nu)_{\nu\in\mathcal{F}}$ a sequence in $\ell^p_m(\mathcal{F})$. If $(\Lambda_n)_{n\geq 1}$ is any lower realization associated with the monotone enveloppe $\mathbf{c}$, then for any $q > p$

$$\Big(\sum_{\nu\notin\Lambda_n} |c_\nu|^q\Big)^{1/q} \leq \Big(\sum_{\nu\notin\Lambda_n} |\mathbf{c}_\nu|^q\Big)^{1/q} \leq \|(c_\nu)\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s_{p,q}}, s_{p,q} := \frac{1}{p} - \frac{1}{q} \tag{1.5.7}$$

For sequences that are in $\ell^p_m(\mathcal{F})$ which is stronger than being in $\ell^p(\mathcal{F})$, it is therefore possible to obtain the same convergence rate as with best n-term approximations in (1.2.13) with the sets Λ_n being lower sets, however with a larger constant $\|(c_\nu)\|_{\ell^p_m(\mathcal{F})}$.

1.5.2 Sparse Taylor approximations in lower sets

We now return to the approximation of the solution map u by Taylor series associated with lower sets. The sequence $(h_\nu)_{\nu\in\mathcal{F}}$ of estimates on Taylor coefficients given in (1.4.19) is monotone decreasing and ℓ^p-summable under the assumptions of Theorem 1.4.5, therefore we have the following

Theorem 1.5.5

Under the uniform ellipticity assumption and if the sequence $b := (\|\psi_j\|_{L^\infty})_{j\geq 1}$ *belongs to* $\ell^p(\mathbb{N})$ *for some* $0 < p < 1$, *then the sequences* $(h_\nu)_{\nu\in\mathcal{F}}$ *and* $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ *belong to* $\ell^p_m(\mathcal{F})$.

In view of this theorem and Lemma 1.5.4, there exists a sequence $(\Lambda_n^{T^*})_{n\geq 1}$ of nested lower sets with $\#(\Lambda_n^{T^*}) = n$ and

$$\Big\|u - \sum_{\nu\in\Lambda_n^{T^*}} t_\nu y^\nu\Big\|_{\mathcal{V}_\infty} \leq \sum_{\nu\notin\Lambda_n^{T^*}} \|t_\nu\|_V \leq \|(\|t_\nu\|_V)\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s}, \qquad s := \frac{1}{p} - 1. \quad (1.5.8)$$

Lower sets for approximation of the map u by Taylor series were introduced in [22] for practical reasons. Indeed, from the recursive formulas (1.3.11), we have that the Taylor coefficients can be computed using the variational problems

$$\int_D \overline{a}\nabla t_\nu \nabla w = -\sum_{j:\nu_j\neq 0}\int_D \psi_j \nabla t_{\nu-e_j}\nabla w, \qquad w\in V. \quad (1.5.9)$$

In order to compute the coefficient t_ν, it is then necessary to know all the coefficient $t_{\nu-e_j}$ for j the actives coordinates of ν which justifies the use of n-term approximations based on nested lower sets.

We should note that in practice the Taylor coefficients are not known in advance, therefore the lower sets $\Lambda_n^{T^*}$ are not known as well. They are of mere theoretical for benchmarking interest in our analysis. Practical construction using adaptive algorithms are presented in Chapter 3.

One way to construct good n-term approximations based on lower sets is thus by using Algorithm 1.5.2 that build a lower realization $(\Lambda_n^h)_{n\geq 1}$ associated with $(h_\nu)_{\nu\in\mathcal{F}}$ which is monotone decreasing. However the sequence $(h_\nu)_{\nu\in\mathcal{F}}$ is given by a double minimization problem and is unlikely to have an explicit closed formula. Even in the simple case of disjoints supports, the sequence $(h_\nu)_{\nu\in\mathcal{F}}$ given in (1.4.21) does not have an explicit formula.

In order to build good computable lower sets, one should rather rely on computable upper bound of the sequence $(h_\nu)_{\nu\in\mathcal{F}}$. For example, the sequence $(q_\nu = (\rho(\nu)^{-\nu}))_{\nu\in\mathcal{F}}$ with the particular $\{r/2\}$-admissible sequences $\rho(\nu)$ introduced in the proof of Theorem

1.4.5 is explicit and satisfies

$$h_\nu \le 2\frac{\|f\|_{V^*}}{r}q_\nu, \qquad \nu \in \mathcal{F}. \tag{1.5.10}$$

The sequence $(q_\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$. However there is no guarantee that this sequence is monotone decreasing. As discussed further in the previous section and in Chapter 3, a finer tuning of the sequence $\rho(\nu)$ can yield a monotone decreasing sequence $(q_\nu)_{\nu\in\mathcal{F}}$.

1.6 Approximation of the solution map with Jacobi polynomials

The solution map u is uniformly bounded over U, therefore it belongs to $\mathcal{V}_\infty \subset \mathcal{V}_2$. This implies that the sum of the Legendre series in (1.2.12) converges unconditionally in $\mathcal{V}_2$ towards u. Moreover, under the assumptions of Theroem 1.3.6, the sequences $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ so that u can be approximated by truncated Legendre series in the least square and uniform sense with the rates in (1.2.15) and (1.2.16) respectively.

The summability result in Theroem 1.3.6 is based on the estimates (1.3.19) which were obtained by the real variable arguments. Better estimates were obtained in [33] using the holomorphy of the solution map u. Namely, it was proved that

$$\|v_\nu\|_V \le \|u_\nu\|_V \le \inf_{0<\delta<r}\Big\{\frac{\|f\|_{V^*}}{\delta}\inf\{\rho^{-\nu}\gamma_\nu(\rho) : \rho \text{ is } \delta\text{-admissible}, \rho_j > 1\}\Big\}, \tag{1.6.1}$$

where

$$\gamma_\nu(\rho) := \prod_{j:\nu_j\neq 0}\phi(\rho_j)(2\nu_j+1), \quad \phi(t) := \frac{\pi t}{2(t-1)}. \tag{1.6.2}$$

The above estimates are larger than the estimates h_ν obtained for Taylor coefficients (1.4.19) due to the presence of the quantities $\gamma_\nu(\rho)$ that are greater than 1. However, the ℓ^p-summability result is unchanged. The estimates were used in [33] in order to establish the following result.

Theorem 1.6.1

Under the uniform ellipticity assumption $\mathbf{UEA}(r, R)$ *and if the sequence* $b := (\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathbb{N})$ *for some* $p < 1$*, then the sequences* $(\|u_\nu\|)_{\nu\in\mathcal{F}}$ *and* $(\|v_\nu\|)_{\nu\in\mathcal{F}}$ *belong to* $\ell^p(\mathcal{F})$*.*

The proof of the previous theorem is similar to that of Theorem 1.4.5. The authors in [33] constructed $\{r/2\}$-admissible sequences $(\rho(\nu))_{\nu\in\mathcal{F}}$ such that $(\rho(\nu)^{-\nu}\gamma_\nu(\rho(\nu)))_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$ proving the ℓ^p-summability of the estimates in (1.6.1).

Unlike the Taylor case, there is no guarantee that the sequence of the estimates (1.6.1) is monotone decreasing. Therefore, the approach of [33] does not discuss if this sequence is in $\ell^p_m(\mathcal{F})$. Similar to Taylor coefficients, we are interested in this stronger type of summability for algorithmic purpose, which are further discussed in Chapter 4 and 5. In this section we will prove this type of summability for Legendre coefficients. Our approach is not exclusive to Legendre expansion. We will present it in the general framework of Jacobi polynomials.

For $\alpha, \beta > -1$, we introduce the Jacobi weight function defined by

$$w_{\alpha,\beta}(t) := (1-t)^\alpha(1+t)^\beta, \qquad t \in [-1,1], \tag{1.6.3}$$

and denote $\varrho_{\alpha,\beta}$ the probability density associated with $w_{\alpha,\beta}$, i.e.

$$\varrho_{\alpha,\beta} := \frac{w_{\alpha,\beta}}{W_{\alpha,\beta}}, \qquad W_{\alpha,\beta} := \int_{-1}^{1} w_{\alpha,\beta}(t)dt. \tag{1.6.4}$$

We denote by $(L_n^{\alpha,\beta})_{n\geq 0}$ the family of univariate Jacobi polynomials associated with $\varrho_{\alpha,\beta}$ which is an orthonormal basis of polynomials in $L^2([-1,1], d\varrho_{\alpha,\beta})$. We note that the constant polynomial $L_0^{\alpha,\beta}$ is equal to 1 because $\varrho_{\alpha,\beta}$ is a probability measure. In order to lighten the notation, we do not give a specific notation for Jacobi polynomials normalized with the supremum over $[-1,1]$ equal to 1. We refer to them when needed with $\frac{L_n^{\alpha,\beta}}{\|L_n^{\alpha,\beta}\|_{L^\infty}}$.

We introduce the tensorized Jacobi polynomials $(L_\nu^{\alpha,\beta})_{\nu\in\mathcal{F}}$ defined by

$$L_\nu^{\alpha,\beta}(y) := \prod_{j\geq 1} L_{\nu_j}^{\alpha,\beta}(y_j), \qquad y = (y_j)_{j\geq 1} \in U. \tag{1.6.5}$$

The family $(L_\nu^{\alpha,\beta})_{\nu\in\mathcal{F}}$ is an orthonormal basis of $L^2(U, d\varrho)$ where $\varrho := \otimes_{j\geq 1}\varrho_{\alpha,\beta}$ denote here the tensorized Jacobi measure. Since the solution map u belongs to $\mathcal{V}_\infty$ then it also belongs to $L^2(U, V, d\varrho)$ and therefore can be expanded according to

$$u = \sum_{\nu\in\mathcal{F}} v_\nu^{\alpha,\beta} L_\nu^{\alpha,\beta}, \qquad v_\nu^{\alpha,\beta} := \int_U u(y)L_\nu^{\alpha,\beta}(y)d\varrho(y). \tag{1.6.6}$$

For $s > 1$, we introduce the Bernstein ellipse in the complex plane

$$\mathcal{E}_s := \Big\{ \frac{w + w^{-1}}{2} : |w| = s \Big\}. \tag{1.6.7}$$

This ellipse has foci 1 and -1 and semi axes of length $\frac{s+s^{-1}}{2}$ and $\frac{s-s^{-1}}{2}$. The ellipse $\mathcal{E}_s$ concentrates near the real interval $[-1, 1]$ when s is close to 1 and grows wider as s increases. It does not contain the unit disc as long as $\frac{s-s^{-1}}{2} < 1$, that is for s in the range $s \in]1, s^*[$ with $s^* = 1 + \sqrt{5}/2$. Also, the convex hull of $\mathcal{E}_s$ is strictly included in the disk $\{|\xi| \leq s\}$.

Given $\rho := (\rho_j)_{j\geq 1}$ a sequence of numbers in $]1, +\infty[$, we denote

$$\mathcal{E}_\rho := \otimes_{j\geq 1} \mathcal{E}_{\rho_j}, \tag{1.6.8}$$

the tensorized poly-ellipse associated with ρ. The convex hull of the poly-ellipse $\mathcal{E}_\rho$ is strictly contained in the polydisc $\mathcal{U}_\rho$ defined in (1.4.13). The following result relates the decay of Jacobi coefficients and the holomorphy of the map u.

Theorem 1.6.2

Let $\rho = (\rho_j)_{j\geq 1}$ be a sequence of real numbers in $]1, +\infty[$. If the solution map u is uniformly bounded by $C_\rho > 0$ over the interior of $\mathcal{E}_\rho$ and holomorphic on a domain $\otimes_{j\geq 1}\mathcal{O}_{\rho_j}$ with $\mathcal{O}_{\rho_j}$ an open neighbourhood of the convex hull of $\mathcal{E}_{\rho_j}$ for every $j \geq 1$, then

$$\|v_\nu^{\alpha,\beta}\|_V \leq C_\rho \prod_{j\geq 1:\nu_j\neq 0} (\nu_j + 1)\varphi(\rho_j)\rho_j^{-\nu_j}, \tag{1.6.9}$$

with the convention that the product is equal to 1 when $\nu = 0_{\mathcal{F}}$ and where $\varphi(t) := \frac{2t}{(t-1)}$ for any $t > 1$.

Proof: In the case $\nu = 0$, the estimate (1.6.9) is immediate since $\varrho(U) = 1$ implies

$$\|v_0^{\alpha,\beta}\|_V = \Big\| \int_U u(y) d\varrho(y) \Big\|_V \leq \sup_{y\in U} \|u(y)\|_V \leq \sup_{z\in\mathcal{E}_\rho} \|u(z)\|_V \leq C_\rho$$

We assume now $\nu \neq 0$ and fixed. Without loss of generality, we assume that ν is supported in $\{1, \ldots, J\}$ for some $J \geq 1$, i.e. $\nu_j \neq 0$ for $1 \leq j \leq J$ and $\nu_j = 0$ for $j \geq J + 1$. This can always be achieved by reordering the basis $(\psi_j)_{j\geq 1}$. We write the variable y as $y = (y_1, \ldots, y_J, y')$ where $y' := (y_{J+1}, y_{J+2}, \ldots) \in U$ and rewrite the coefficient $v_\nu^{\alpha,\beta}$ defined in (1.6.6) as

$$v_\nu^{\alpha,\beta} = \int_U w_\nu(y') d\varrho(y'),$$

where

$$w_\nu(y') := \int_{[-1,1]^J} u(y_1, \ldots, y_J, y') \left(\prod_{j=1}^J L_{\nu_j}^{\alpha,\beta}(y_j) \right) d\varrho_{\alpha,\beta}(y_1) \ldots d\varrho_{\alpha,\beta}(y_J).$$

Since $\varrho(U)=1$, then $\|v_\nu^{\alpha,\beta}\|_V \leq \sup_{y'\in U}\|w_\nu(y')\|_V$. We propose to show that for every $y'\in U$, the quantity $\|w_\nu(y')\|_V$ is smaller than the right side in the inequality (1.6.9) above.

We fix $y'\in U$. By the holomorphy assumption, the map $(z_1,\dots,z_J)\mapsto u(z_1,\dots,z_J,y')$ is holomorphic on $\otimes_{j=1}^J\mathcal{O}_{\rho_j}$. Every domain $\mathcal{O}_{\rho_j}$ is a neighboured of the interior of $\mathcal{E}_{\rho_j}$, therefore applying inductively Cauchy's integral formula in each ellipse $\mathcal{E}_{\rho_j}$ in the variable z_j for $j=1,...,J$, we obtain that for any $(y_1,\dots,y_J)\in[-1,1]^J$

$$u(y_1,\dots,y_J,y') = \frac{1}{(2\pi i)^J}\int\limits_{\mathcal{E}_{\rho,J}} \frac{u(z_1,\dots,z_J,y')}{(y_1-z_1)\dots(y_J-z_J)}dz_1\dots dz_J,$$

with $\mathcal{E}_{\rho,J}=\otimes_{j=1}^J\mathcal{E}_{\rho_j}$. Multiplying by $\prod_{j=1}^J L_{\nu_j}^{\alpha,\beta}(y_j)$, integrating over $[-1,1]^J$ with respect to $d\varrho_{\alpha,\beta}(y_1)\dots d\varrho_{\alpha,\beta}(y_J)$ and interchanging integration orders, we obtain

$$w_\nu(y') = \frac{1}{(2\pi i)^J}\int\limits_{\mathcal{E}_{\rho,J}} u(z_1,\dots,z_J,y')\Big(\prod_{j=1}^J\int_{-1}^1 \frac{L_{\nu_j}^{\alpha,\beta}(y_j)}{z_j-y_j}d\varrho_{\alpha,\beta}(y_j)\Big)dz_1\dots dz_J,$$

Since $(z_1,\dots,z_J,y')$ is in the interior of $\mathcal{E}_\rho$ for any $(z_1,\dots,z_J)\in\mathcal{E}_{\rho,J}$, then using uniform boundedness assumption, we deduce

$$\|w_\nu(y')\|_V \leq C_\rho\Big(\prod_{j=1}^J\rho_j\Big)\Big(\prod_{j=1}^J\sup_{\xi\in\mathcal{E}_{\rho_j}}\Big|\int_{-1}^1\frac{L_{\nu_j}^{\alpha,\beta}(t)}{\xi-t}d\varrho_{\alpha,\beta}(t)\Big|\Big),$$

where we have used the fact that each of the ellipses $\mathcal{E}_{\rho_j}$ has perimeter of length less or equal to $2\pi\rho_j$. We complete the proof using the corollary A.3.2 of the appendix, in which we prove that

$$\sup_{\xi\in\mathcal{E}_s}\Big|\int_{-1}^1\frac{L_n^{\alpha,\beta}(t)}{\xi-t}d\varrho_{\alpha,\beta}(t)\Big| \leq 2(n+1)\frac{s^{-n}}{s-1}, \qquad s>1.$$

■

By inspection of the previous proof, we note that the bound (1.6.9) for $\nu\in\mathcal{F}-\{0\}$ stays valid if the sequence ρ satisfies $\rho_j=1$ for any $j\geq 1$ such that $\nu_j=0$. An immediate implication of the previous theorem is the estimation of Jacobi coefficients for the solution map u. Given $\delta\in]0,r[$ and $\rho=(\rho_j)_{j\geq1}$ az δ-admissible sequence with $\rho_j>1$ for any $j\geq1$, we have seen that u is holomorphic and uniformly bounded by $\frac{\|f\|_{V^*}}{\delta}$ over $\mathcal{U}_\rho$. Since every $\mathcal{E}_{\rho_j}$ is contained in the open disc $\{|\xi|<\rho_j\}$, then the bound (1.6.9) holds with $C_\rho=\frac{\|f\|_{V^*}}{\delta}$. We have then,

Theorem 1.6.3

Under the uniform ellipticity assumption **UEA**(r,R), *for any* $\nu\in\mathcal{F}$

$$\|v_\nu^{\alpha,\beta}\|_V \leq \Big(\prod_{j\geq1}(\nu_j+1)\Big)\inf_{0<\delta<r}\Big\{\frac{\|f\|_{V^*}}{\delta}\inf\{\rho^{-\nu}\prod_{j\geq1:\nu_j\neq0}\varphi(\rho_j) : \rho \text{ is } \delta\text{-admissible}, \rho_j>1\}\Big\} \tag{1.6.10}$$

We now turn to the summability of the previous estimates. As with Legendre polynomials, we are also interested in studying the summability of the Jacobi coefficients $(u_\nu^{\alpha,\beta})_{\nu\in\mathcal{F}}$ associated with Jacobi polynomials normalized in $L^\infty(U)$. Since these Jacobi coefficients are given by

$$u_\nu^{\alpha,\beta} = v_\nu^{\alpha,\beta}\|L_\nu^{\alpha,\beta}\|_{L^\infty} = v_\nu^{\alpha,\beta}\prod_{j\geq 1}\|L_n^{\alpha,\beta}\|_{L^\infty}, \tag{1.6.11}$$

and since, as shown in (A.2.4) in the appendix, the supremum norm of univariate Jacobi polynomials is controlled by

$$\|L_n^{\alpha,\beta}\|_{L^\infty} \leq Cn^\gamma, \quad \gamma = \max\left\{\frac{2\alpha+1}{2}, \frac{2\beta+1}{2}, 0\right\},\ C = C(\alpha,\beta), \tag{1.6.12}$$

their study amounts to the study of the summability of the estimates in (1.6.14) deteriorated by multi-dimensional algebraic factors of the type $C_\nu(\theta)$ defined in 1.3.26 and encountered in Theorem 1.3.3. Studying this more general case turn our to be useful also for the analysis in chapters 2 and 5.

We introduce a new notation for admissibility that we will adopt in the following chapters.

Definition 1.6.4

Given a sequence $b := (b_j)_{j\geq 1}$ of strictly positive real numbers and $\varepsilon > 0$, we say that the sequence $(\rho_j)_{j\geq 1}$ is (b,ε)-admissible if and only if $\rho_j \geq 1$ for any $j \geq 1$ and

$$\sum_{j\geq 1}(\rho_j - 1)b_j < \varepsilon \tag{1.6.13}$$

In the present setting, denoting by b the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$, we observe that for $\varepsilon \in]0,1[$ the (b,ε)-admissibility implies the δ-admissibility introduced in (1.4.11) with $\delta = r - \varepsilon$. This equality being satisfied, the inequality (1.6.14) implies

$$\|v_\nu^{\alpha,\beta}\|_V \leq \frac{\|f\|_{V^*}}{r-\varepsilon}\Big(\prod_{j\geq 1}(\nu_j+1)\Big)g_\nu, \qquad \nu\in\mathcal{F}, \tag{1.6.14}$$

where we have introduced the sequence $(g_\nu)_{\nu\in\mathcal{F}}$ defined by $g_{0_\mathcal{F}} = 1$ and

$$g_\nu := \inf\left\{\rho^{-\nu}\prod_{j\geq 1:\nu_j\neq 0}\varphi(\rho_j) : \rho \text{ is } (b,\varepsilon)\text{-admissible}, \rho_j > 1\right\}. \tag{1.6.15}$$

The following theorem implies the summability result for the sequences $(\|v_\nu^{\alpha,\beta}\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu^{\alpha,\beta}\|_V)_{\nu\in\mathcal{F}}$.

Theorem 1.6.5

Let $\varepsilon > 0$ be arbitrary and $(g_\nu)_{\nu\in\mathcal{F}}$ as above. If the sequence b belongs $\ell^p(\mathbb{N})$ for some $0 < p < 1$, then for any $C, \theta > 0$ the sequence $\Big(C_\nu(\theta)\ g_\nu\Big)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$.

Remark 1.6.6

Before giving the proof, let us observe that this result is a significant improvement over Theorem 1.4.5 which was used in order to prove the ℓ^p summability for the Taylor coefficients. Indeed, the quantities g_ν are generally larger than h_ν for two reasons. On the one hand, they contain the factors $C_\theta(\nu)\prod_{j\geq 1:\nu_j\neq 0}\varphi(\rho_j)$ in the quantity to be infimized. On the other hand, when $\varepsilon > 0$ is small enough, the property of (b,ε)-admissibility used in the definition of g_ν will imply the property of δ-admissibility used in the definition of h_ν and therefore the infimum is taken over a smaller set of sequences ρ. In addition, we prove an even stronger result by using $\ell^p_m(\mathcal{F})$ in place of $\ell^p(\mathcal{F})$.

Proof: Let $B > 0$ be arbitrary but fixed. Since $0 < p < 1$, the sequence b belongs to $\ell^1(\mathbb{N})$, let then $J \geq 1$ be an integer such that

$$\sum_{j>J} |b_j| \leq \frac{\varepsilon}{4B}$$

We write $\mathbb{N} = E \cup F$ with $E := \{1 \leq j \leq J\}$ and $F := \{j \geq J+1\}$ and introduce for $\nu = (\nu_j)_{j\geq 1} \in \mathcal{F}$ the notations $\nu_E := (\nu_1, \ldots, \nu_J) \in \mathbb{N}^J$ and $\nu_F := (\nu_{J+1}, \nu_{J+2}, \ldots) \in \mathcal{F}$. Let $\nu \in \mathcal{F} - \{0\}$ fixed. We introduce the sequence $\rho(\nu) := (\rho_j)_{j\geq 1}$ that depends on ν according to

$$\rho_j := \kappa \quad \text{for} \quad j \in E \quad \text{and} \quad \rho_j := \kappa + B + \frac{\varepsilon}{2|b_j|}\frac{\nu_j}{|\nu_F|+1} \quad \text{for} \quad j \in F,$$

where $\kappa = 1 + \frac{\varepsilon}{4\|(b_j)\|_{\ell^1}}$. We have

$$\sum_{j\geq 1}(\rho_j - 1)b_j \leq (\kappa - 1)\sum_{j\geq 1} b_j + B\sum_{j>J} b_j + \frac{\varepsilon}{2}\sum_{j>J}\frac{\nu_j}{|\nu_F|+1} \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon$$

The sequence $\rho(\nu)$ is then (b,ε)-admissible. Therefore from (1.6.15)

$$C_\nu(\theta)g_\nu \leq \prod_{j\geq 1:\nu_j\neq 0} C\nu_j^\theta\varphi(\rho_j)\rho_j^{-\nu_j} = \prod_{j\leq J:\nu_j\neq 0} C\nu_j^\theta\varphi(\kappa)\kappa^{-\nu_j}\prod_{j>J:\nu_j\neq 0} C\nu_j^\theta\varphi(\rho_j)\rho_j^{-\nu_j}.$$

We introduce the notation $C_\kappa = \varphi(\kappa) > 1$. We have $\phi(\rho_j) \leq C_\kappa$ for any $j \geq 1$ because φ is a monotone decreasing function and $\rho_j \geq \kappa$ for any $j \geq 1$. This combined with the crude bounds $CC_\kappa n^\theta\kappa^{-n} \leq C_1\kappa^{-\frac{n}{2}}$ and $CC_\kappa n^\theta \leq C_2^n$ for any $n \geq 1$ for some constants $C_1, C_2 > 1$ and with $\rho_j \geq B + \frac{\varepsilon}{2|b_j|}\frac{\nu_j}{|\nu_F|+1}$ for any $j > J$, implies

$$C_\nu(\theta)g_\nu \leq C_1^J q_\nu,$$

with $q_\nu := q_E(\nu)q_F(\nu)$ and

$$q_E(\nu) := \prod_{j\le J} \kappa^{-\nu_j/2} \quad \text{and} \quad q_F(\nu) := \prod_{j>J:\nu_j\neq 0} C_2^{\nu_j}\Big(B + \frac{\varepsilon}{2|b_j|}\frac{\nu_j}{1+|\nu_F|}\Big)^{-\nu_j}.$$

In view of $g_{0_\mathcal{F}} = 1$ and $C_1 > 1$, the previous bound is valid for $\nu = 0_\mathcal{F}$ with $q_E(0_\mathcal{F}) = q_F(0_\mathcal{F}) = 1$. By following the same line as the proof of Theorem 1.4.5, we obtain by choosing J sufficiently large, that the sequence $(q_\nu)_{\nu\in\mathcal{F}}$ is proved to belong to $\ell^p(\mathcal{F})$.

In order to prove that $(g_\nu)_{\nu\in\mathcal{F}}$ is in $\ell^p_m(\mathcal{F})$, we propose to show that, if B is chosen sufficiently large, the sequence $(q_\nu)_{\nu\in\mathcal{F}}$ can be made monotone decreasing. We define $e_j := (0,\dots,0,1,0,\dots)$ the Kronecker sequence with 1 at position j. It is easily checked that $q_{\nu+e_j} = \kappa^{-1/2}q_\nu \le q_\nu$ for any $\nu \in \mathcal{F}$ and any $j \le J$. Now, we consider $j \ge J$. We have

$$q_{e_j} = C_2\Big(B + \frac{\varepsilon}{4|b_j|}\Big)^{-1} \le 1 = q_{0_\mathcal{F}},$$

when $B > C_2$. For $\nu \in \mathcal{F} - \{0\}$, we have

$$\frac{q_{\nu+e_j}}{q_\nu} = \frac{C_2}{B + \frac{\varepsilon}{2b_j}\frac{\nu_j+1}{2+|\nu_F|}}\Big(\frac{B + \frac{\varepsilon}{2b_j}\frac{\nu_j}{1+|\nu_F|}}{B + \frac{\varepsilon}{2b_j}\frac{\nu_j+1}{2+|\nu_F|}}\Big)^{\nu_j} \prod_{k>J:k\neq j,\ \nu_k\neq 0}\Big(\frac{B + \frac{\varepsilon}{2b_k}\frac{\nu_k}{1+|\nu_F|}}{B + \frac{\varepsilon}{2b_k}\frac{\nu_k}{2+|\nu_F|}}\Big)^{\nu_k}.$$

The term in the middle is smaller than $\Big(\frac{B+\frac{\varepsilon}{2b_j}\frac{\nu_j}{1+|\nu_F|}}{B+\frac{\varepsilon}{2b_j}\frac{\nu_j}{2+|\nu_F|}}\Big)^{\nu_j}$. This combined with $\frac{B+A_1}{B+A_2} \le \frac{A_1}{A_2}$ for any $A_1 \ge A_2 > 0$, implies

$$\frac{q_{\nu+e_j}}{q_\nu} \le \frac{C_2}{B}\prod_{k>J:\nu_k\neq 0}\Big(\frac{B + \frac{\varepsilon}{2b_j}\frac{\nu_k}{1+|\nu_F|}}{B + \frac{\varepsilon}{2b_j}\frac{\nu_k}{2+|\nu_F|}}\Big)^{\nu_k} \le \frac{C_2}{B}\prod_{k>J:\nu_k\neq 0}\Big(\frac{2+|\nu_F|}{1+|\nu_F|}\Big)^{\nu_k} = \frac{C_2}{B}\Big(1+\frac{1}{1+|\nu_F|}\Big)^{|\nu_F|} \le \frac{eC_2}{B}.$$

where we have used $(1+\frac{1}{x+1})^x \le e$ for any $x \ge 1$. Therefore $q_{\nu+e_j} \le q_\nu$ if $B > eC_2$. We deduce that, for B sufficiently large, the sequence $(q_\nu)_{\nu\in\mathcal{F}}$ is monotone decreasing. Therefore $(q_\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$ and so does $(g_\nu)_{\nu\in\mathcal{F}}$ ■

In view of the estimates (1.6.14) of Jacobi coefficients and of (1.6.12), the following result is an immediate consequence of the previous Theorem.

Theorem 1.6.7

Under the uniform ellipticity assumption **UEA**(r, R) *and if* $b = (\|\psi_j\|_{L^\infty(D)})_{j\ge 1}$ *belongs to* $\ell^p(\mathbb{N})$ *for some* $0 < p < 1$, *the sequences* $(\|v^{\alpha,\beta}_\nu\|_V)_{\nu\in\mathcal{F}}$ *and* $(\|u^{\alpha,\beta}_\nu\|_V)_{\nu\in\mathcal{F}}$ *belongs to* $\ell^p_m(\mathcal{F})$.

Using Lemmas 1.2.1 and 1.5.4, under the assumptions of the previous theorem, we are able to translate the conclusion of the above theorem in terms of convergence rates for sparse Jacobi approximations. First, using only the ℓ^p summability in a similar fashion to the approximations with Legendre series (1.2.15) and (1.2.15), we find that

if Λ_n^L and Λ_n^P are index sets associated to the n largest terms in $(\|v_\nu^{\alpha,\beta}\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu^{\alpha,\beta}\|_V)_{\nu\in\mathcal{F}}$ respectively, then

$$\Big\|u - \sum_{\nu\in\Lambda_n^L} v_\nu^{\alpha,\beta} L_\nu^{\alpha,\beta}\Big\|_{L^2(U,V,d\varrho)} \le \|(v_\nu)\|_{\ell^p(\mathcal{F})}(n+1)^{-s^*}, \quad s^* := \frac{1}{p} - \frac{1}{2}\,. \tag{1.6.16}$$

and

$$\Big\|u - \sum_{\nu\in\Lambda_n^P} v_\nu^{\alpha,\beta} L_\nu^{\alpha,\beta}\Big\|_{\mathcal{V}_\infty} \le \|(u_\nu)\|_{\ell^p(\mathcal{F})}(n+1)^{-s}, \quad s := \frac{1}{p} - 1, \tag{1.6.17}$$

Second, using the ℓ_m^p summability, we have that if $(\Lambda_n^L)_{n\ge1}$ and $(\Lambda_n^P)_{n\ge1}$ are lower realizations associated with the monotone envelopes of the sequences $(\|v_\nu^{\alpha,\beta}\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu^{\alpha,\beta}\|_V)_{\nu\in\mathcal{F}}$ respectively, then

$$\Big\|u - \sum_{\nu\in\Lambda_n^L} v_\nu^{\alpha,\beta} L_\nu^{\alpha,\beta}\Big\|_{L^2(U,V,d\varrho)} \le \|(v_\nu)\|_{\ell_m^p(\mathcal{F})}(n+1)^{-s^*}, \quad s^* := \frac{1}{p} - \frac{1}{2}\,. \tag{1.6.18}$$

and

$$\Big\|u - \sum_{\nu\in\Lambda_n^P} v_\nu^{\alpha,\beta} L_\nu^{\alpha,\beta}\Big\|_{\mathcal{V}_\infty} \le \|(u_\nu)\|_{\ell_m^p(\mathcal{F})}(n+1)^{-s}, \quad s := \frac{1}{p} - 1, \tag{1.6.19}$$

In consequence, the n-term truncated Jacobi series provide approximations to the solution map u in $\mathcal{V}_\infty$ with similar convergence rates as the Taylor series and provide approximations with better decay rate in $L^2(U,V,d\varrho)$ (that coincides with $\mathcal{V}_2$ in the Legendre case).

1.7 Conclusion

In this chapter, we have presented the paradigm of [34, 33] concerned with the study of the elliptic model with diffusion coefficient depending affinely in the parameter vector $y = (y_j)_{j\ge1}$. The analysis can be generalized in a straightforward manner to many other classes of parametric PDE. For example, for a separable Hilbert space V, consider the equation

$$Au = f, \tag{1.7.1}$$

where f belongs to V^* and where A is an operator from V to V^*. We assume that A depends affinely on $y \in U$ according to

$$A = A(y) = A_0 + \sum_{j\ge0} y_j \Psi_j \tag{1.7.2}$$

where A_0 and the Ψ_j are operators from V to V^*. We assume that A is uniformly continuous and uniformly coercive, i.e.

$$\langle A(y)v, v\rangle \ge r\|v\|_V^2 \quad \text{and} \quad |\langle A(y)v, w\rangle| \le R\|v\|_V\|w\|_V, \quad y\in U, \quad v,w\in V \tag{1.7.3}$$

for some $0 < r \leq R < \infty$. We introduce the sequence $b = (b_j)_{j\geq 0}$ defined by

$$b_j := \|\Psi_j\|_{V\to V^*} = \sup_{\substack{v,w\in V\\ \|v\|=\|w\|=1}} |\langle \Psi_j v, w\rangle|. \tag{1.7.4}$$

Using the exact same arguments as in the previous sections, it can be shown that $b \in \ell^p(\mathbb{N})$ for some $p < 1$ implies that the solution map $y \in U \mapsto u(y) \in V$ can be approximated by its Taylor and Legendre series with algebraic rate $(n+1)^{-s}$ and $(n+1)^{-s^*}$ and that the truncated series can be localized to lower sets.

The paradigm can also be applied directly to the parametric parabolic equation

$$\partial_t u - \operatorname{div}(a\nabla u) = f, \quad \text{in} \quad [0,T]\times D, \tag{1.7.5}$$

with

$$u|_{\partial D} = 0 \quad \text{for} \quad 0 < t < T \quad \text{and} \quad u|_{t=0} = u_0 \in V\ , \quad \text{for any} \quad y \in U\ . \tag{1.7.6}$$

where f and a are as in the linear elliptic model studied throughout this chapter. Here the solution space is

$$V := L^2(0,T;H^1_0(D)) \cap H^1(0,T;H^{-1}(D)). \tag{1.7.7}$$

Again, using the exact same arguments as in the previous sections, it can be shown that under the assumption $(\|\psi_j\|_{L^\infty})_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, the solution map u can be approximated by it Taylor series and Legendre series with the rates $(n+1)^{-s}$ and $(n+1)^{-s^*}$.

We can push further the generalization of the paradigm for many other types of parametric PDEs. For example, it is of interest to inspect the complex analysis arguments in §1.4 for possible generalization on the assumption on f. In particular, we observe that we can have f parametric in which case it is a map

$$f : U \mapsto V^* \tag{1.7.8}$$

and assume that f can be extended by holomorphy to all the domains $\mathcal{U}_\rho$ for any δ-admissible sequence ρ, see (1.4.11), with f uniformly bounded on these domains

$$M := \sup_{\rho\ \ \delta-\text{admissible}} \sup_{z\in\mathcal{U}_\rho} \|f(z)\|_{V^*} < \infty. \tag{1.7.9}$$

Under such assumptions, the holomorphy and uniform boundedness of the solution u map on domains $\mathcal{U}_\rho$ for δ-admissible sequences ρ, with δ fixed, can be established. Using this, one derives the same bounds (1.4.19) for Taylor coefficients with M instead of $\|f\|_{V*}$ and the summability analysis is unchanged. The generalization over f can of course be also considered for models (1.7.1) and (1.7.5).

The paradigm introduced so far therefore applies to various classes of parametric models. However, our analysis is strongly tied to (i) the linearity of the models and (ii) the affine dependence of the operators on y. It is not as easily applicable for models that do not satisfy such prescriptions, even very simple one, such as the same elliptic equation

$$-\mathrm{div}(a\nabla u) = f \tag{1.7.10}$$

where a would have a non affine form such as

$$a = \exp\Big(1 + (\sum_{j\geq 1} y_j\psi_j)^2\Big), \tag{1.7.11}$$

or the semi-linear model

$$u^3 - \mathrm{div}(a\nabla u) = f, \tag{1.7.12}$$

even in the case where a is affine in y. In the next chapter we use the key points of the paradigm, namely uniform holomorphy, uniform boundedness and best n-term approximation, in order propose a more general framework allowing the treatment of problematic models such as the above.

Chapter 2

A framework for general parametric PDEs

Contents

2.1 Introduction

In this chapter, we investigate the numerical approximation of more general class of parametric PDEs than the elliptic equation treated in Chapter 1. For such equations, we adopt the abstract formulation from the introduction, thus considering the general form

$$\mathcal{D}(u, y) = 0, \tag{2.1.1}$$

where $u \mapsto \mathcal{D}(u, y)$ is a partial differential linear or nonlinear operator that depends on an infinite parameter vector $y = (y_j)_{j\geq 1} \in U = [-1, 1]^{\mathbb{N}}$. Assuming that for any $y \in U$, the above problem is well posed in a certain Banach space V, we again introduce the *solution map*

$$y \in U \mapsto u(y) \in V \ . \tag{2.1.2}$$

A typical setting for high dimensional parametric PDEs occurs for problems which are parametrized by a *function* h varying over a certain class, according to

$$\mathcal{P}(u, h) = 0, \tag{2.1.3}$$

where $\mathcal{P}$ is a given partial differential linear or nonlinear operator. The function h may for example used to describe or parametrize (i) a spatially variable diffusion property of a material as in Chapter 1 (ii) a flux function in a transport problem (iii) a forcing term such as the right hand side f in Chapter 1 (iv) the geometry of the physical domain. Using a given basis $(\psi_j)_{j\geq 1}$ for expanding h and rescaling the corresponding coefficients, we may write

$$h = h(y) := \sum_{j\geq 1} y_j \psi_j, \qquad y \in U, \tag{2.1.4}$$

which yields the parametric model (2.1.1), with

$$\mathcal{D}(u, y) := \mathcal{P}(u, h(y)) = \mathcal{P}\Big(u, \sum_{j\geq 1} y_j \psi_j\Big), \tag{2.1.5}$$

and where the number of variables is now countably infinite, that is $d = \infty$, or very large if the above expansion has been truncated with high accuracy. This situation was shortly described in the introduction (1.10) and (1.11) where typically the infinite series (2.1.4) of h, as a random field, results from its Karhunen-Loève expansion.

As for the elliptic model in Chapter 1, the large number d of variables of the solution map u is a serious obstruction, because of the curse of dimensionality. Numerical approximation of u requires then non-standard discretization tools and a description of the smoothness of this map which differs from the classical description in terms of C^m spaces. A key idea is to introduce more subtle models which reflect the *anisotropy* of this map in the sense that it has a weaker or smoother dependence on certain variables than others. Intuitively this is due to the fact that the convergence of the series (2.1.4) for all $y \in U$ should typically be reflected by a certain form of decay in the size of ψ_j as $j \to +\infty$, resulting in weaker dependence on the corresponding variables y_j. As a consequence the discretization tools should also reflect this anisotropy.

The effectiveness of the previously described paradigm was demonstrated in Chapter 1 for the elliptic model. We have indeed seen in Theorem 1.4.5 that if the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$ has a decay that we characterized by $\ell^p(\mathbb{N})$-summability for some $0 < p < 1$, then the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ of the norms of Taylor coefficients inherits the

same decay, which implies the approximation of u by its truncated Taylor series with an algebraic rate $(n+1)^{-s}$ with $s = \frac{1}{p} - 1$ in the uniform sense as in (1.2.14). The rate is proved despite the fact that $d = \infty$, showing that one can in principle overcome the curse of dimensionality in the approximation of u by a proper choice of sparse polynomial spaces.

The proof of Theorem 1.4.5 is based on the analysis of the anisotropic holomorphy of the solution map, in the sense of extending it to the complex domain and making a fine study of its region of holomorphy in several complex variables. Unfortunately this latter aspect is heavily tied to the affine dependence of the diffusion coefficient a with respect to the parameter y as in (1.1.2) and to the linear nature of the elliptic equation (1.1.1).

Many practically relevant parametric PDEs are nonlinear and depend on the parameters y in a non-affine manner. The objective of the present chapter is to propose a general strategy in order to derive similar polynomial approximation results for such PDEs. Here are a few examples, among many others, that can be treated by the approach introduced in this chapter:

(i) Operator equations such as (1.1.1), with non-affine, yet holomorphic, dependence in y of the diffusion coefficients and such that the problem is well posed uniformly in $y \in U$. Typical instances are

$$a(x,y) := \overline{a} + \Big(\sum_{j\geq 1} y_j \psi_j\Big)^2, \tag{2.1.6}$$

with $\overline{a}$ a strictly positive function which satisfies $\overline{a}(x) \geq r > 0$ for any $x \in D$, or

$$a(x,y) = \exp\Big(\sum_{j\geq 1} y_j \psi_j\Big), \quad y \in U \tag{2.1.7}$$

so that the solution $u(y)$ of (1.1.1) is uniquely defined in $V = H^1_0(D)$.

(ii) Linear parabolic evolution equations with spatial operators as in (i). Specifically, for a coefficient a as in (i), we consider in the Gel'fand evolution triple $X \subset Y \simeq Y^* \subset X^*$ the parabolic problem

$$\partial_t u - \operatorname{div}(a\nabla u) - f = 0 \quad \text{in} \quad (0,T)\times D, \tag{2.1.8}$$

where $f \in L^2(0,T;X^*)$, with initial and boundary conditions

$$u|_{\partial D} = 0 \quad \text{for} \quad 0 < t < T\,, \quad \text{and} \quad u|_{t=0} = u_0 \in Y\,, \quad \text{for any} \quad y \in U\,. \tag{2.1.9}$$

We consider here $X = H^1_0(D)$ and $Y := L^2(D)$. A solution space (see [41]) for the parametric PDE is

$$V := L^2(0,T;X) \cap H^1(0,T;X^*). \tag{2.1.10}$$

Boundary conditions, other than homogeneous Dirichlet, can be accommodated with other choices of the space X.

(iii) Nonlinear operator equations, with analytic dependence of $\mathcal{D}$ on u and on y and such that the problem is uniformly well posed in $y \in U$. One typical instance is the monotone, elliptic problem

$$u^{2q+1} - \operatorname{div}(a\nabla u) - f = 0, \tag{2.1.11}$$

which is set on a bounded liptschiz physical domain $D \subset \mathbb{R}^m$ of dimension $m \geq 1$ and with homogeneous Dirichlet boundary conditions on ∂D and right-hand side $f \in H^{-1}(D)$, where a depends on y as in (1.1.2) and where $q \geq 0$ is an integer such that $q < \frac{m}{m-2}$. These conditions ensure existence and uniqueness of the solution $u(y)$ in $V = H^1_0(D)$, for every $y \in U$, by the theory of monotone operators (see Chapter 6 of [72]).

(iv) Operator equations on domains whose shape depends on a parameter sequence y. As a simple example, consider the Laplace equation

$$-\Delta v = f, \tag{2.1.12}$$

with homogeneous Dirichlet boundary conditions set on a physical domain $D(y) \subset \mathbb{R}^2$ that depends on y in the following manner

$$D(y) := \{(x_1, x_2) \ : \ 0 \leq x_1 \leq 1, \ 0 \leq x_2 \leq \phi(x_1, y)\}, \tag{2.1.13}$$

where $\phi(t, y) := \overline{\phi} + \sum_{j\geq 1} y_j \psi_j(t)$ satisfies a condition of the same type as (1.1.3) ensuring that the boundary of $D(y)$ is not self-intersecting. Using the map $\Phi(y)(x_1, x_2) := (x_1, x_2\phi(x_1))$ one can transport back the solution $v(y) \in H^1_0(D(y))$ into the reference domain $D = [0, 1]^2$ according to $u(y) := v(y) \circ \Phi(y) \in H^1_0(D)$. The functions $u(y)$ are solutions to an elliptic PDE set on D with diffusion coefficient and source term that both depend on the parameter sequence y in a holomorphical, but non-affine manner.

The strategy developed in Chapter 1 for proving Theorem 1.4.5 for the model equation (1.1.1) with coefficients given by (1.1.2) does not carry over for the above problems. In fact, this theorem will generally fail to hold, in the sense that in the previously described models, we may assume $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$ and yet the Taylor series of u does not converge in $L^\infty(U, V)$. This is due to the fact that, for the above models, the solution map does not generally admit an holomorphic extension in a neighbourhood of the whole unit polydisc

$$\mathcal{U} := \otimes_{j\geq 1}\{|z_j| \leq 1\} \ . \tag{2.1.14}$$

As a simple example, consider model (i) or (ii) with $a(x,y) = 1+by_1^2$, as a particular case of (2.1.6) where b is a constant strictly larger than 1. In this case, a has no spacial variability and the solution map u has an explicit formula,

$$u(y) := \frac{u(0)}{1+by_1^2}, \qquad y \in U. \tag{2.1.15}$$

The holomorphy in the first variable on an open disc $\{|z_1| < \rho_1\}$ may then holds only if $\rho_1 \leq b^{-1/2} < 1$. A more elaborate inspection of models (iii) and (iv) reveals similar problems. A different approach is therefore needed for the construction and the convergence analysis of sparse polynomial approximation.

An alternative to Taylor series are the Legendre series or more generally Jacobi series, which were also investigated in Chapter 1. We have seen in particular with Theorem 1.6.2 and its implications that weaker assumptions on the analyticity of u, for instance in domains that do not necessarily contain the unit disc $\mathcal{U}$, yields the convergence of Legendre series toward u with the rate $(n+1)^{-s}$ as with Taylor series. The rate is improved in the mean square sense into $(n+1)^{-s^*}$, where $s^* := \frac{1}{p} - \frac{1}{2}$.

In the present chapter, we show that a large variety of models, including in particular (i)-(ii)-(iii)-(iv), can be treated using Legendre series. As in the proof of Theorem 1.6.2, the key argument will consist in extending u analytically to neighbourhoods of domains of the type

$$\mathcal{E}_\rho := \otimes_{j\geq 0}\mathcal{E}_{\rho_j}, \tag{2.1.16}$$

where $\mathcal{E}_s$ is the Bernstein ellipse (1.6.7) introduced in §1.6. As in Chapter 1, we use the admissible range of radii ρ_j which reflect the anisotropy of the problem, in order to establish ℓ^p-summability results on Legendre coefficients.

In §2.2, we introduce a property referred to as (p,ε)-holomorphy assumption, or shortly $\mathbf{HA}(p,\varepsilon)$, that describes the domains of holomorphy of the solution map as poly-ellipses of the form (2.1.16). We show that this property yields upper bounds for the V-norms of Legendre coefficients $\|u_\nu\|_V$ which allows us to derive $\ell^p(\mathcal{F})$ summability results as in Chapter 1.

In §2.3, we introduce a first framework that allows us to establish the validity of $\mathbf{HA}(p,\varepsilon)$ for various classes of linear parametric PDEs. More precisely, this framework applies to linear variational problems in Hilbert spaces, based on the inf-sup (LBB) theory.

In §2.4, we introduce a second framework for establishing $\mathbf{HA}(p,\varepsilon)$ for parametric PDEs which are not necessarily linear. For instance, PDE based on semi-linear on non-linear differential operators $\mathcal{D}$. This framework generalizes the first one and deals with operators $\mathcal{D}$ which have a smooth dependence on u and holomorphic dependence on y and is based on the implicit function theorem in complex Banach spaces.

Finally, we discuss in §2.5 the application of the two introduced frameworks to the

previously described models (i) to (iv). We show that (iii) and (iv) can be treated using the second framework and that both frameworks may be used to treat (i) and (ii).

2.2 The (p, ϵ)-holomorphy and implications

We consider sparse approximations constructed by tensorized Legendre series

$$u(y) = \sum_{\nu \in \mathcal{F}} v_\nu L_\nu(y) = \sum_{\nu \in \mathcal{F}} u_\nu P_\nu(y), \tag{2.2.1}$$

where $(L_\nu)_{\nu \in \mathcal{F}}$ and $(P_\nu)_{\nu \in \mathcal{F}}$ are the tensorized Legendre polynomials that we introduced in Chapter 1, formula (1.2.6). We recall that the family $(L_\nu)_{\nu \in \mathcal{F}}$ form an orthonormal basis of the space $\mathcal{V}_2 := L^2(U, V, d\varrho)$ of square integrable, V-valued map with respect to the uniform product probability measure ϱ defined in (1.2.8). The expansion (2.2.1) is then justified whenever $u \in \mathcal{V}_2$ or more simply u belongs to $\mathcal{V}_\infty \subset L^\infty(U, V)$ the space of functions defined everywhere in U and uniformly bounded in V.

As we have seen in chapter 1, the use of Legendre series instead of Taylor series allows us to obtain similar sparse approximation results under weaker assumptions on the domains of holomorphic extension of the solution map u. In particular, our analysis relies on holomorphic extensions of u over domains of the type $\mathcal{E}_\rho = \otimes_{j \geq 1} \mathcal{E}_{\rho_j}$ where $\mathcal{E}_s$ for $s > 1$ denote the ellipse defined in (1.6.7). Our approach consists then in assuming minimal assumptions on the operator $\mathcal{D}$ enabling to extend u to such domains.

We recall the notation of (b, ε)-admissibility that we introduced in Chapter 1. Given $b := (b_j)_{j \geq 1}$ a sequence of strictly positive real numbers and $\varepsilon > 0$, we denote $\mathcal{A}_{\varepsilon, b}$ the set of (b, ε)-admissible sequences associated with b and ε, as in (1.6.13). In other words

$$\mathcal{A}_{\varepsilon, b} := \Big\{ \rho := (\rho_j)_{j \geq 1} \in]1, +\infty[^{\mathbb{N}} : \sum_{j \geq 1} (\rho_j - 1) b_j < \varepsilon \Big\}, \tag{2.2.2}$$

In light of the previous discussion and the paradigm developed in Section 1.6 of Chapter 1, the minimal assumption to be supposed on the operator $\mathcal{D}$ is introduced in the following definition.

Definition 2.2.1

For $\varepsilon > 0$ and $0 < p < 1$, we say that $\mathcal{D}$ satisfies the (p, ε)-holomorphy assumption, denoted **HA**(p, ε)*, if and only if*

(i) For every $y \in U$, there exists a unique solution $u(y) \in V$ of the problem (2.1.1) and the map $y \in U \mapsto u(y) \in V$ is uniformly bounded, i.e.

$$\sup_{y \in U} \|u(y)\|_V \leq C_0 \,, \tag{2.2.3}$$

for some finite constant $C_0 > 0$.

(ii) There exists a positive sequence $(b_j)_{j\geq 1} \in \ell^p(\mathbb{N})$ *and a constant* $C_\varepsilon > 0$ *such that for any sequence* ρ *in* $\mathcal{A}_{\varepsilon,b}$*, the map* u *admits a complex extension* $z \mapsto u(z)$ *that is holomorphic with respect to each variable* z_j *on a domain of the form* $\mathcal{O}_\rho := \otimes_{j\geq 1}\mathcal{O}_{\rho_j}$*, with every* $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ *is an open set containing* $\mathcal{E}_{\rho_j}$*. This extension is bounded on* $\mathcal{E}_\rho := \otimes_{j\geq 1}\mathcal{E}_{\rho_j}$*, according to*

$$\sup_{z\in\mathcal{E}_\rho} \|u(z)\|_X \leq C_\varepsilon. \tag{2.2.4}$$

The first assumption justifies the existence of the Legendre coefficients and the equality (2.2.1) in $\mathcal{V}_2$. Moreover, in view of Thereom 1.6.2, the second assumption implies,

$$\|v_\nu\|_V \leq \Big(\prod_{j\geq 1}(\nu_j+1)\Big)\, C_\varepsilon \inf_{\rho\in\mathcal{A}_{\varepsilon,b}} \Big\{\rho^{-\nu} \prod_{j\geq 1:\nu_j\neq 0} \varphi(\rho_j)\Big\}, \qquad \nu\in\mathcal{F} \tag{2.2.5}$$

with the infimum is equal to 1 when $\nu = 0_\mathcal{F}$ and the function φ is defined for $t > 1$ by $\varphi(t) := \frac{2t}{(t-1)}$. Let us note that in Theorem 1.6.2, we have worked with the $H^1_0(D)$-norm of the coefficients v_ν, while here we work with the V-norm. The inspection of the proof of the theorem shows that the space in which u take its values is irrelevant. The sequence on the right hand of (2.2.5) is of the form $(C_\varepsilon C_\nu(\theta) g_\nu)_{\nu\in\mathcal{F}}$ where the g_ν defined with the same notation in (1.6.15). Therefore using Theorem 1.6.5, we are able to deduce the following,

Theorem 2.2.2

If the differential operator $\mathcal{D}$ *is such that* $\mathbf{HA}(p,\varepsilon)$ *holds for some* $0 < p < 1$ *and* $\varepsilon > 0$*, then the sequences* $(\|u_\nu\|_X)_{\nu\in\mathcal{F}}$ *and* $(\|v_\nu\|_X)_{\nu\in\mathcal{F}}$ *belong to* $\ell^p_m(\mathcal{F})$*, and*

$$u(y) = \sum_{\nu\in\mathcal{F}} v_\nu L_\nu = \sum_{\nu\in\mathcal{F}} u_\nu P_\nu, \tag{2.2.6}$$

holds in the sense of unconditional convergence in $\mathcal{V}_\infty$.

Using Stechkin lemma 1.2.1, we can translate the conclusion of the above theorem in terms of convergence rates for sparse Legendre approximations as in (1.6.18) and (1.6.19): if $(\Lambda^P_n)_{n\geq 1}$ and $(\Lambda^L_n)_{n\geq 1}$ are sequence of nested lower sets corresponding respectively to the n largest terms in the monotone envelopes $\mathbf{u} := (\mathbf{u}_\nu)_{\nu\in\mathcal{F}}$ and $\mathbf{v} := (\mathbf{v}_\nu)_{\nu\in\mathcal{F}}$ of the sequences $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$, then

$$\Big\|u - \sum_{\nu\in\Lambda^P_n} u_\nu P_\nu\Big\|_{\mathcal{V}_\infty} \leq \|(\|u_\nu\|_X)\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s}, \quad s := \frac{1}{p} - 1, \tag{2.2.7}$$

and

$$\Big\|u - \sum_{\nu\in\Lambda_n^L} v_\nu L_\nu\Big\|_{\mathcal{V}_2} \le \|(\|v_\nu\|_X)\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s^*}, \quad s^* := \frac{1}{p} - \frac{1}{2}\,. \tag{2.2.8}$$

In consequence, whenever a parametric PDE is described by an operator $\mathcal{D}$ that satisfies the (p,ε)-holomorphy assumption, then its solution map u can be approximated by the n-term truncated Legendre series in the uniform and mean square senses with algebraic rates. The interest of having lower sets is useful for the construction of computable approximations, for instance using interpolation, sparse grids and least square, as we shall see in Chapters 5-6.

The notion of (p,ε)-holomorphy is merely intended for formalization purposes. Indeed, Definition 2.2.1 does not give concrete assumptions on $\mathcal{D}$ but rather on the solution map u. We therefore need frameworks where one consider specific assumptions on $\mathcal{D}$ that can imply the (p,ε)-holomorphy and be verified for models such as (i)-(ii)-(iii)-(iv). We propose two general frameworks in this direction. These frameworks are rather simple and can be applied for many potential models of parametric PDEs.

In the case of models (i), (ii) and (iii), we verify $\mathbf{HA}(p,\varepsilon)$, using $b_j := \|\psi_j\|_{L^\infty(D)}$ and under the assumption that $(\|\psi_j\|_{L^\infty(D)})_{j\ge1}$ belongs to $\ell^p(\mathbb{N})$. In the case of model (iv), we establish the validity of $\mathbf{HA}(p,\varepsilon)$ using $b_j := \|\psi_j\|_{L^\infty(D)} + \|\psi_j'\|_{L^\infty(D)}$, and therefore under the additional assumption that $(\|\psi_j'\|_{L^\infty(D)})_{j\ge1}$ belongs to $\ell^p(\mathbb{N})$.

2.3 The linear variational framework

The first framework is concerned with the parametric PDE that has the general variational form: for any $y\in U$, the function $u(y)\in V$ is solution of

$$B(u(y),v,y) = F(v,y), \qquad v\in W, \tag{2.3.1}$$

where V, W are Hilbert spaces over $\mathbb{C}$ and where, for every fixed $y\in U$, the maps $(u,v)\mapsto B(u,v,y)$ and $v\mapsto F(v,y)$ are continuous sesquilinear and antilinear forms on $V\times W$ and on W respectively. In this setting, the operator $\mathcal{D}$ of (2.1.1) is defined from $V\times U$ into the antidual W^* of W, according to

$$\mathcal{D}(u,y) := B(u,\cdot,y) - F(\cdot,y)\,. \tag{2.3.2}$$

In many practical instances, the two spaces V and W coincide, however $V\neq W$ is relevant for the treatment of parabolic evolution problems. We use the same notations B and F to denote the corresponding maps from U into the spaces of sesquilinear and antilinear continuous forms on $V\times W$ and on W, respectively, defined by

$$B(y)(v,w) := B(v,w,y) \quad\text{and}\quad F(y)(w) := F(w,y), \qquad y\in U,\ v\in V,\ w\in W. \tag{2.3.3}$$

We propose a framework that makes possible the treatment of the parametric variational problem (2.3.1) using minimal assumption on B and F. As for the linear model in Chapter 1, we introduce first an assumption allowing the well-posedness of (2.3.1) for any $y \in U$ and any z in the complex domains $\mathcal{O} \in \mathbb{C}^{\mathbb{N}}$ where the solution map u might be extended. This is given in the following generic definition

Definition 2.3.1

Let $\mathcal{O}$ be an open domain of $\mathbb{C}^{\mathbb{N}}$ and assume that B and F can be extended over $\mathcal{O}$. We say that B and F satisfies the uniform continuity and inf-sup assumption if and only if there exist constants $0 < r \leq R < \infty$ and $0 < M < \infty$ not depending on z such that for any $z \in \mathcal{O}$

$$\sup_{w \in W - \{0\}} \frac{|F(w,z)|}{\|w\|_W} \leq M, \qquad \sup_{\substack{v \in V - \{0\} \\ w \in W - \{0\}}} \frac{|B(v,w,z)|}{\|v\|_V \|w\|_W} \leq R, \tag{2.3.4}$$

and

$$\inf_{v \in V - \{0\}} \sup_{w \in W - \{0\}} \frac{|B(v,w,z)|}{\|v\|_V \|w\|_W} \geq r, \qquad \inf_{w \in W - \{0\}} \sup_{v \in V - \{0\}} \frac{|B(v,w,z)|}{\|v\|_V \|w\|_W} \geq r. \tag{2.3.5}$$

First, note that by assuming the uniform continuity and inf-sup inequalities above for the real domain U with constants r_0, R_0 and M_0 insures, using a standard functional analytic argument similar to the proof of the Lax-Milgram lemma, that the parametric problem (2.3.1) is well posed in V for any $y \in U$ and that the solution map $y \mapsto u(y)$ is uniformly bounded, with

$$\sup_{y \in U} \|u(y)\|_V \leq \frac{M_0}{r_0}. \tag{2.3.6}$$

Moreover, under only theses assumptions with U, we can study the regularity of the map $y \in U \mapsto u(y)$ using real variable arguments similarly to the affine linear model in Section 1.3 of chapter 1. We dot not embark in this direction, but rather study the additional minimal assumptions on B and F that convey to $\mathcal{D}$ the (p, ε)-holomorphy

Definition 2.3.2

For $\varepsilon > 0$ and $0 < p < 1$, we say that F and B satisfies the (p, ε)-holomorphy assumption if and only if there exists a positive sequence $(b_j)_{j \geq 1} \in \ell^p(\mathbb{N})$, and two constants $0 < r \leq R < \infty$ and a constant $M < \infty$ such that the following holds:

(i) For any sequence $\rho := (\rho_j)_{j \geq 1}$ in $\mathcal{A}_{b,\varepsilon}$, the maps B and F admit extensions that are holomorphic with respect to every variable z_j on a set of the form $\mathcal{O}_\rho = \otimes_{j \geq 1} \mathcal{O}_{\rho_j}$, where every $\mathcal{O}_{\rho_j} \subset \mathbb{C}$ is an open set containing $\mathcal{E}_{\rho_j}$.

(ii) These extensions satisfy the uniform continuity and inf-sup assumptions of

the definition 2.3.1 over the domains $\mathcal{O}_\rho$ with the constants r, R and M.

The following result shows that the validity of $\mathbf{HA}(p,\varepsilon)$ expressing the analytic behavior of the solution map $y \mapsto u(y)$ follows from the same analytic behavior of the maps B and F expressed in the previous definition.

Theorem 2.3.3

For $\varepsilon > 0$ and $0 < p < 1$, assume that B and F satisfy the analytic assumption of Definition 2.3.2 with a sequence b and constants r, R and M. Then, the corresponding operator $\mathcal{D}$ defined as in (2.3.2) satisfies the assumption $\mathbf{HA}(p,\varepsilon)$ with the same p and ε and with the same sequence b.

Proof: Let p, ε, b, $\rho := (\rho_j)_{j\geq 1}$ and $\mathcal{O}_\rho$ be as in the assumptions of Theorem 2.3.3. First, using the continuity and inf-sup conditions (2.3.4) and (2.3.5), a standard functional analytic argument similar to the proof of the Lax-Milgram lemma, shows that for any $z \in \mathcal{O}_\rho$, the variational problem

$$\mathcal{D}(u,z) := B(z)(u,\cdot) - L(z)(\cdot) = 0 \quad \text{in} \quad W^* \tag{2.3.7}$$

is well posed in V and its solution $u(z)$ is bounded by $\frac{M}{r}$ where r and M are the same as in conditions (2.3.4) and (2.3.5). Accordingly, the solution map $z \in \mathcal{O}_\rho \mapsto u(z) \in V$ is well-defined and uniformly bounded in $\mathcal{O}_\rho$ with

$$\sup_{z\in\mathcal{O}_\rho} \|u(z)\|_X \leq \frac{M}{r}, \tag{2.3.8}$$

In order to complete the proof of Theorem 2.3.3, we only need to prove that u is holomorphic in $\mathcal{O}_\rho$ with respect to each variable z_j. We first observe that u is continuous on $\mathcal{O}_\rho$: indeed, for $z, \tilde{z} \in \mathcal{O}_\rho$, we have from the equations $\mathcal{D}(u(z),z) = 0$ and $\mathcal{D}(u(\tilde{z}),\tilde{z}) = 0$ in W^* that

$$B(z)\Big(u(\tilde{z}) - u(z), v\Big) = -\Big(B(\tilde{z}) - B(z)\Big)(u(\tilde{z}),v) + \Big(F(\tilde{z}) - F(z)\Big)(v), \quad v \in W\,. \tag{2.3.9}$$

Therefore, taking $v = u(\tilde{z}) - u(z)$ and using the continuity and inf-sup conditions (2.3.4) and (2.3.5), we obtain

$$r\|u(\tilde{z})-u(z)\|_V^2 \leq \|B(\tilde{z})-B(z)\|_{\mathcal{L}(V\times W,\mathbb{C})}\|u(\tilde{z})\|_V\|u(\tilde{z})-u(z)\|_V + \|F(\tilde{z})-F(z)\|_{W^*}\|u(\tilde{z})-u(z)\|_V,$$

which combined with (2.3.8) implies

$$\|u(\tilde{z}) - u(z)\|_V \leq \frac{1}{r}\Big(\|B(\tilde{z}) - B(z)\|_{\mathcal{L}(V\times W,\mathbb{C})}\frac{M}{r} + \|F(\tilde{z}) - F(z)\|_{W^*}\Big).$$

The holomorphy of B and F implies then the continuity of u. Now, let $z \in \mathcal{O}_\rho$, $j \geq 1$ and $\delta \in \mathbb{C}$ such that $z + \delta e_j \in \mathcal{O}_\rho$, where e_j is the j-th Kronecker sequence in $\mathbb{C}^{\mathbb{N}}$ and

introduce $w_\delta = \frac{1}{\delta}(u(z+\delta e_j) - u(z))$. Taking $z+\delta e_j$ in place of $\tilde{z}$ in (2.3.9) , we obtain that for every $v \in W$

$$B(z)(w_\delta, v) = -\frac{B(z+\delta e_j) - B(z)}{\delta}(u(z+\delta e_j), v) + \frac{F(z+\delta e_j) - F(z)}{\delta}(v), \qquad v \in W \tag{2.3.10}$$

By the holomorphic dependence of B and L on z,

$$\left\| \frac{F(z+\delta e_j) - F(z)}{\delta} - \frac{\partial F}{\partial z_j}(z) \right\|_{W^*} = o_\delta(1) \quad \text{and} \quad \left\| \frac{B(z+\delta e_j) - B(z)}{\delta} - \frac{\partial B}{\partial z_j}(z) \right\|_{\mathcal{L}(V\times W,\mathbb{C})} = o_\delta(1),$$

where we use the generic notation $o_\delta(1)$ for a positive quantity that tends to 0 as $\mathbb{C} \ni \delta \to 0$. Using again (2.3.8) to bound the functions $u(z+\delta e_j)$ in (2.3.10) for any δ such that $z+\delta e_j \in \mathcal{O}_\rho$, we infer that for any $v \in Y$

$$\left| B(z)(w_\delta, v) - \frac{\partial F}{\partial z_j}(z)(v) + \frac{\partial B}{\partial z_j}(z)(u(z+\delta e_j), v) \right| = \|v\|_W o_\delta(1).$$

This, combined with the continuous dependence of u on z, implies

$$\left\| B(z)(w_\delta, \cdot) - \frac{\partial F}{\partial z_j}(z)(\cdot) + \frac{\partial B}{\partial z_j}(z)(u(z), \cdot) \right\|_{W^*} = o_\delta(1).$$

Finally, we denote $w_0 \in V$ the unique solution of the variational problem

$$B(z)(w_0, \cdot) = \frac{\partial F}{\partial z_j}(z)(\cdot) - \frac{\partial B}{\partial z_j}(z)(u(z), \cdot), \quad \text{in} \quad W^*.$$

The existence of w_0 is insured by the uniform inf-sup condition on B, the continuity of $\frac{\partial F}{\partial z_j}$ and $\frac{\partial B}{\partial z_j}$, and the inequality (2.3.8) for bounding $u(z)$ in V. We have that

$$\|B(z)(w_\delta - w_0, \cdot)\|_{W^*} = o_\delta(1) \ .$$

The inf-sup condition in (2.3.5) then implies $\|w_\delta - w_0\|_V \to 0$. This shows that the map $z \mapsto u(z)$ from $\mathbb{C}$ to V admits a partial complex derivative $\frac{\partial u}{\partial z_j}(z) \in V$ with respect to the complex extension z_j of each coordinate variable y_j. In addition, this derivative is the unique solution of the variational problem

$$B(z)\Big(\frac{\partial u}{\partial z_j}(z), v\Big) = \frac{\partial F}{\partial z_j}(z)(v) - \frac{\partial B}{\partial z_j}(z)(u(z), v), \quad v \in W \ . \tag{2.3.11}$$

The proof of the holomorphy of u with respect to every variable on $\mathcal{O}_\rho$ is then complete. ■

We notice that in the previous proof, the partial derivative $\partial_{e_j} u(z)$ satisfies the same variational problem the the one by $u(z)$ but a with linear form defined by

$$v \in W \mapsto \partial_{e_j} F(z)(v) - \partial_{e_j} B(z)(u(z), v). \tag{2.3.12}$$

Since this form is also holomorphic over the considered domains $\mathcal{O}_\rho$, we can reiterate the arguments of the previous proof and show that u admits partial derivatives to any

order $\nu \in \mathcal{F}$ at any $z \in \mathcal{O}_\rho$. Such derivatives can be obtained by deriving formally the variational formula satisfied by $u(z)$ and using Leibniz derivation formula. In particular, for any $z \in \mathcal{O}_\rho$, the partial derivative $\partial_\nu u(z)$ is the unique solution of the following variational problem

$$B(z)\Big(\partial_\nu u(z), v\Big) = \partial_\nu F(z)(v) - \sum_{\mu<\nu} \frac{\nu!}{\mu!(\nu-\mu)!} \partial_{\nu-\mu} B(z)(\partial_\mu u(z), v), \quad v \in W. \quad (2.3.13)$$

Let us also note that the inspection of the previous proof shows that u inherits the same holomorphy regions where B and F are holomorphic. In particular, if B and F satisfies the (p,ε)-holomorphy assumption given in Definition 2.3.2 in the polydisks $\mathcal{U}_\rho$ instead of the ellipses $\mathcal{E}_\rho$, we can show using the arguments of Chapter 1 that the solution map can be approximated by its Taylor series at 0 and the series $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ of Taylor coefficients is in $\ell^p(\mathcal{F})$. The Taylor coefficients $t_\nu := \frac{\partial_\nu u(0)}{\nu!}$ can be computed using the recursive formula (2.3.13) by taking $z = 0$ in which case one has

$$B(0)(t_\nu, v) = \frac{\partial_\nu F(0)}{\nu!}(v) - \sum_{\mu<\nu} \frac{\partial_{\nu-\mu} B(0)}{(\nu-\mu)!}(t_\mu, v), \quad v \in W. \quad (2.3.14)$$

Remark 2.3.4

Inspection of the proof of Theorem 2.3.3 reveals that it remains valid verbatim when V and W are reflexive Banach spaces.

2.4 The implicit function theorem framework

Our second framework is concerned with parametric PDEs of the form (2.1.3). The operator $\mathcal{D}$ depends on the parameter $y \in U$ through the functions

$$h(y) = \sum_{j\geq 1} y_j \psi_j \quad (2.4.1)$$

where the functions ψ_j belong to some Banach space L over $\mathbb{C}$. We assume that the expansion converges in L for all $y \in U$. The operator $\mathcal{D}$ depends on y according to

$$\mathcal{D}(u, y) = \mathcal{P}\Big(u, h(y)\Big), \quad (2.4.2)$$

where $\mathcal{P}$ is a linear or nonlinear operator defined from the product of the two Banach spaces V and L over $\mathbb{C}$ into a third Banach space W over $\mathbb{C}$. In the particular case of the elliptic model in Chapter 1, we have $V = H_0^1(D)$, $L = L^\infty(D)$ and $W = H^{-1}(D)$. We introduce the set

$$h(U) = \Big\{ h(y) \ : \ y \in U \Big\} \subset L \, . \quad (2.4.3)$$

We set $b := (b_j)_{j\geq 1}$ with $b_j := \|\psi_j\|_L$. As for the elliptic model, we propose to use this sequence to establish the $\mathbf{HA}(p,\varepsilon)$-holomorphy of the operator $\mathcal{D}$. The validity of $\mathbf{HA}(p,\varepsilon)$ is ensured provided that $(b_j)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$ and that $\mathcal{P}$ satisfies certain smoothness properties, in addition to the well-posedness of the problem (2.1.3) over $h(U)$. These properties are given in Theorem (2.4.3) below.

Before giving Theorem 2.4.3, we give two simple, yet useful observations that reveal the key points in the present framework. The first observation is concerned with the topology of the set $h(U)$ introduced in (2.4.3). The second observation is concerned with the open neighborhood $\mathcal{O}_s$ for the complex ellipse $\mathcal{E}_s$ in which we propose to establish the holomorphy of the map u.

Lemma 2.4.1

Assume that the the sequence $(\|\psi_j\|_L)_{j\geq 1}$ belongs to $\ell^1(\mathbb{N})$. Then $h(U)$ is compact in L.

Proof: Let $(h_n)_{n\geq 1}$ be a sequence in $h(U)$. Since $(\|\psi_j\|_L)_{j\geq 1} \in \ell^1(\mathbb{N})$, the sequence $(h_n)_{n\geq 1}$ is bounded in L. Each h_n is of the form $h_n = \sum_{j\geq 1} y_{n,j}\psi_j$. Using a Cantor diagonal argument, we infer that there exists $y = (y_j)_{j\geq 1} \in U$ such that

$$\lim_{n\to+\infty} y_{\sigma(n),j} = y_j, \quad j \geq 1, \tag{2.4.4}$$

where $(\sigma(n))_{n\geq 1}$ is a monotone sequence of positive integers. Defining $h := \sum_{j\geq 1} y_j\psi_j \in h(U)$, we may write for any $k \geq 1$,

$$\|h_{\sigma(n)} - h\|_L \leq \Big\| \sum_{j=1}^{k} (y_j - y_{\sigma(n),j})\psi_j \Big\|_L + 2 \sum_{j\geq k+1} \|\psi_j\|_L. \tag{2.4.5}$$

It follows that $h_{\sigma(n)}$ converges towards h in L and therefore $h(U)$ is compact. ∎

Lemma 2.4.2

Let $s > 1$ and introduce the set in $\mathbb{C}$

$$\mathcal{O}_s := \bigcup_{t\in[-1,1]} \{\xi \in \mathbb{C} : |\xi - t| < s - 1\} = \{\xi \in \mathbb{C} \ : \ \mathrm{dist}(\xi, [-1,1]) < s - 1\}\ . \tag{2.4.6}$$

The set $\mathcal{O}_s$ is an open neighbourhood of the convex hull of $\mathcal{E}_s$.

Proof: The set $\mathcal{O}_s$ is open by construction, therefore it is sufficient to prove that $\mathcal{E}_s \subset \mathcal{O}_s$. Since the ellipse $\mathcal{E}_s$ has half-axes $\frac{s+s^{-1}}{2}$ and $\frac{s-s^{-1}}{2}$ and foci ± 1, then for any $\xi \in \partial\mathcal{E}_s$ we have

(i) If $\Re(\xi) \in [-1,1]$, then since $|\Im(\xi)| \leq \frac{s-s^{-1}}{2} < s - 1$, we have $|\xi - \Re(\xi)| < s - 1$.

(ii) If $\Re(\xi) > 1$ then $|\xi + 1| > 2$, but since $|\xi - 1| + |\xi + 1| = s + s^{-1}$, we have $|\xi - 1| < s + s^{-1} - 2 < s - 1$.

(iii) If $\Re(\xi) < -1$, then by symmetry with (ii), we have $|\xi + 1| < s - 1$.

This shows that in the three cases $|\xi - t| < s - 1$ for some $t \in [-1, 1]$ and completes the proof. ■

We now can give the main Theorem of this section. The idea of the framework is to extend the solution map $y \mapsto u(y)$ by holomorphy on complex neighbourhoods of the form $\otimes_{j\geq 1}\{|z_j - y_j| < \epsilon_j(y)\}$ of any $y \in U$, then by a compactness argument show that as y varies in U the radius $\epsilon_j(y)$ stay bounded from below by some ϵ_j, which in view of Lemma (2.4.2) implies that u can be extended to $\mathcal{E}_\rho$ with $\rho := (1 + \epsilon_j)_{j\geq 1}$. The following theorem provide a general setting where the previous arguments apply.

Theorem 2.4.3

Assume that:

- *One has* $(\|\psi_j\|_L)_{j\geq 1} \in \ell^p(\mathbb{N})$ *for some* $0 < p < 1$.
- *The problem* (2.1.3) *is well-posed in* X *for all* $h \in h(U)$.
- *The map* $(u, h) \mapsto \mathcal{P}(u, h)$ *is continuously differentiable from* $V \times L$ *into* W.
- *For every* $h \in h(U)$, *the partial differential* $\frac{\partial \mathcal{P}}{\partial u}(u(h), h)$ *is an isomorphism from* V *onto* W.

Then there exists an $\varepsilon > 0$, *for which* $\mathcal{D}$ *satisfies the assumptions* $\mathbf{HA}(p, \varepsilon)$.

Proof: We consider an arbitrary $y \in U$ and the corresponding $h(y) \in h(U)$. The assumptions of Theorem 2.4.3 say that $\mathcal{P}$ is continuously differentiable as a mapping from $V \times L$ into W, that $\mathcal{P}(u(y), h(y)) = 0$ in W and that the partial differential $\frac{\partial \mathcal{P}}{\partial u}(u(y), h(y))$ is an isomorphism from V onto W. Therefore, by the holomorphic version of the implicit function theorem on complex Banach spaces, see [44, Theorem 10.2.1], there exists an $\varepsilon > 0$, and a mapping G from $\mathring{\mathcal{B}}(h(y), \varepsilon)$ the open ball of L with center $h(y)$ and radius ε into V such that $G(h(y)) = u(y)$ and $\mathcal{P}(G(h), h) = 0$ for any h in $\mathring{\mathcal{B}}(h(y), \varepsilon)$. In addition, the map G is uniformly bounded and holomorphic on $\mathring{\mathcal{B}}(h(y), \varepsilon)$ with

$$dG(h) = -\Big(\frac{\partial \mathcal{P}}{\partial u}(G(h), h)\Big)^{-1} \circ \frac{\partial \mathcal{P}}{\partial h}(G(h), h), \quad h \in \mathring{\mathcal{B}}(h(y), \varepsilon)\,. \tag{2.4.7}$$

Let us note that $\varepsilon = \varepsilon(y)$ depends actually on y. We claim that ε can be made independent of $y \in U$. Since $\bigcup_{y\in U} \mathring{\mathcal{B}}(h(y), \frac{\varepsilon(y)}{2})$ is an infinite open covering of $h(U)$ and since $h(U)$ is compact in L, thanks to Lemma 2.4.1, then there exists a finite sub-cover of $h(U)$, i.e. a finite number M and $y^1, \cdots, y^M$ in U such that

$$h(U) \subset \bigcup_{j=1}^{M} \mathring{\mathcal{B}}\Big(h(y^j), \frac{\varepsilon(y^j)}{2}\Big)\,. \tag{2.4.8}$$

We introduce $\varepsilon := \min_{1\le j\le M} \frac{\varepsilon(y^j)}{2}$. Let $y \in U$ and $h \in L$ such that $\|h - h(y)\|_L < \varepsilon$. According to (2.4.8), $h(y)$ belongs to some $\mathring{\mathcal{B}}(h(y^j), \frac{\varepsilon(y^j)}{2})$, therefore for $j = 1, \dots, M$

$$\|h - h(y^j)\|_L \le \|h - h(y)\|_L + \|h(y) - h(y^j)\|_L < \varepsilon + \frac{\varepsilon(y^j)}{2} \le \frac{\varepsilon(y^j)}{2} + \frac{\varepsilon(y^j)}{2} = \varepsilon(y^j).$$

This shows that $\mathring{\mathcal{B}}(h(y), \varepsilon) \subset \mathring{\mathcal{B}}(h(y^j), \varepsilon(y^j))$ and it implies that

$$h^\varepsilon(U) := \bigcup_{y\in U} \mathring{\mathcal{B}}(h(y), \varepsilon) \subset \bigcup_{j=1}^{M} \mathring{\mathcal{B}}(h(y^j), \varepsilon(y^j)) \ . \tag{2.4.9}$$

In particular the map G is well defined and is continuously differentiable as a mapping from $h^\varepsilon(U)$ into the complex Banach space V.

To conclude the proof of Theorem 2.4.3, we verify assumption $\mathbf{HA}(p, \varepsilon)$. Let $\rho := (\rho_j)_{j\ge1}$ a sequence of numbers strictly greater than 1 such that $\sum_{j\ge1}(\rho_j - 1)b_j \le \varepsilon$ and $\mathcal{O}_\rho := \otimes_{j\ge1}\mathcal{O}_{\rho_j}$, where for $s > 1$, $\mathcal{O}_s$ is the open domain in $\mathbb{C}$ defined in (2.4.6). For any $z := (z_j)_{j\ge1} \in \mathcal{O}_\rho$, we define $h(z) := \sum_{j\ge1} z_j\psi_j \in L$. If $y = (y_j)_{j\ge1} \in U$ satisfies $|z_j - y_j| < \rho_j - 1$ for every $j \ge 1$, we then have

$$\|h(z)-h(y)\|_L = \left\| \sum_{j\ge1}(z_j-y_j)\psi_j \right\|_L \le \sum_{j\ge1} |z_j-y_j|\|\psi_j\|_L < \sum_{j\ge1}(\rho_j-1)b_j \le \varepsilon, \tag{2.4.10}$$

therefore $h(z) \in h^\varepsilon(U)$ and $G(h(z))$ is well defined. We extend the solution map u on the domain $\mathcal{O}_\rho$ by $u(z) := G(h(z))$. By holomorphy of G on $h^\varepsilon(U)$ and affine dependence of $h(z)$ on z, it follows that

$$z \mapsto h(z) \mapsto u(z) = G(h(z)),$$

is holomorphic with respect to every variable z_j on $\mathcal{O}_\rho$. Moreover

$$\sup_{z\in\mathcal{O}_\rho} \|u(z)\|_X = \sup_{z\in\mathcal{O}_\rho} \|G(h(z))\|_X \le \sup_{h\in h^\varepsilon(U)} \|G(h)\|_X \le \max_{i=1,\dots,M} \sup_{h\in\mathring{\mathcal{B}}(h(y^j),\varepsilon(y^j))} \|G(h)\|_X < \infty \ . \tag{2.4.11}$$

This completes the proof of Theorem 2.4.3. ■

Remark 2.4.4

Inspection of the above proof reveals that we can weaken the assumption in the sense that holomorphy of the map $\mathcal{P}$ is required only over a set of the form $V \times h_\eta(U)$ for some $\eta > 0$ instead of $V \times L$, where $h_\eta(U) := \{h \in L \ : \ \mathrm{dist}_L(h, h(U)) < \eta\}$.

We should note that the previous theorem can also apply in order to prove the (p, ε)-holomorphy of the operator $\mathcal{D}$ associated with the parametric problem (2.3.1) of the first framework. More precisely, if the bilinear form B and the linear form F in (2.3.1) depends on $y \in U$ through $h(y)$, we have the operator

$$\mathcal{P}(u, h) := B(u, \cdot, h) - F(\cdot, h), \quad (u, h) \in V \times L. \tag{2.4.12}$$

We only need to assume that $(\|\psi_j\|_L)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$ and that B and F satisfies the uniform continuity and inf-sup conditions given in Definition 2.3.1 over U (that is for every $h \in h(U)$). Indeed, as we have seen in the previous section, this last assumption implies both the well posedness and the continuous differentiability with respect to h. In addition, the bi-linearity of B implies that $\mathcal{P}$ is continuously differentiable with respect to the variable u with $\frac{\partial \mathcal{P}}{\partial u}(u,h) = B(u,\cdot,h)$. The inf-sup condition over U implies then the forth assumption of Theorem (2.4.3). The assumptions of Theorem (2.4.3) are then complete.

2.5 Application to general models

In this section, we show that the models (i)-(ii)-(iii)-(iv) discussed in the introduction are covered by at least one of the two frameworks of Theorem 2.3.3 or Theorem 2.4.3. Specifically, we check the assumptions of Theorem 2.3.3 for models (i)-(ii)-(iv) and of Theorem 2.4.3 for models (i)-(ii)-(iii).

2.5.1 Model (i): Linear elliptic PDEs with non-affine parametric coefficients

We recall that model (i) is the parametric elliptic diffusion equation (1.1.1) with the typical instances of the diffusion coefficient a

$$a(x,y) := \overline{a}(x) + \Big(\sum_{j\geq 1} y_j \psi_j(x)\Big)^2 \qquad \text{or} \qquad a(x,y) := \exp\Big(\sum_{j\geq 1} y_j \psi_j\Big), \quad x \in D, \quad y \in U. \tag{2.5.1}$$

In both cases, assuming that the well posedness is guaranteed in the Sobolev space $V = H_0^1(D)$, the solution map u is given by: For any $y \in U$, the function $u(y) \in V$ is the unique solution of the variational problem

$$B(u(y), w, y) = F(w, y), \quad w \in V, \tag{2.5.2}$$

where we have defined the sesquilinear and antilinear forms by

$$B(u,w,y) := \int_D a(x,y)\nabla u(x)\overline{\nabla w(x)}dx \qquad \text{and} \qquad F(w,y) := F(w) = \int_D f(x)\overline{w(x)}dx \tag{2.5.3}$$

In order to apply the first framework (Theorem 2.3.3) to the present elliptic models, we need to verify the (p,ε)-holomorphy of B and F given in Definition (2.3.2). Since $f \in V^* = H^{-1}(D)$ does not depends on y, then F can be extended by holomophy to $\mathbb{C}^{\mathbb{N}}$ and it is uniformly continuous with the constant

$$M := \|f\|_{V^*}. \tag{2.5.4}$$

We are left with the study of B which depends on y through the diffusion coefficient a.

Quadratic diffusion coefficient:

For the first example, we assume that $\bar{a}$ and the function ψ_j are in $L := L^\infty(D)$ with $\bar{a}$ bounded from below by some $r_0 > 0$ and the sequence $b := (\|\psi_j\|_L)_{j\geq 1}$ belongs to $\ell^p(\mathbb{N})$ for some $p < 1$. This implies that a satisfies a uniform ellipticity assumption of type (1.1.3) with

$$r_0 := \min_{x\in D} \bar{a}(x) \qquad \text{and} \qquad R_0 := \|\bar{a}\|_L + \|b\|^2_{\ell^1(\mathbb{N})}, \tag{2.5.5}$$

and establishes the well-posedness of (1.1.1) in V for any $y \in U$.

Now, we extend the diffusion coefficient a by holomorphy to complex variables $z = (z_j)_{j\geq 1} \in \mathbb{C}^{\mathbb{N}}$ in the natural way by replacing the y_j by z_j in the above expression (2.5.1). Given any sequence $\rho = (\rho_j)_{j\geq 0}$ of numbers strictly greater than 1, we denote $\mathcal{O}_\rho := \otimes_{j\geq 1}\mathcal{O}_{\rho_j}$, with $\mathcal{O}_s \subset \mathbb{C}$ is the domain defined in (2.4.6). The assumptions of Definition 2.3.2 can be fulfilled with p, M, the domains $\mathcal{O}_\rho$ and the numbers

$$\varepsilon := \sqrt{\frac{r_0}{2}}, \qquad r := \frac{r_0}{2}, \qquad R := R_0 + 2\varepsilon^2 + \|b\|^2_{\ell^1}. \tag{2.5.6}$$

Indeed, given $\rho = (\rho_j)_{j\geq 0}$ a sequence of number strictly greater than 1 with $\sum_{j\geq 1}(\rho_j - 1)b_j \leq \varepsilon$, we have for $x \in D$ and $z \in \mathcal{O}_\rho$

$$\begin{aligned} \Re(a(x,z)) &= \bar{a}(x) + \Big(\sum_{j\geq 1} \Re(z_j)\psi_j(x)\Big)^2 - \Big(\sum_{j\geq 1} \Im(z_j)\psi_j(x)\Big)^2 \\ &\geq r_0 - \Big(\sum_{j\geq 1} |\Im(z_j)|b_j\Big)^2 \geq r_0 - \Big(\sum_{j\geq 1}(\rho_j - 1)b_j\Big)^2 \geq r_0 - \varepsilon^2 = r. \end{aligned} \tag{2.5.7}$$

We have used that for $s > 1$ the domain $\mathcal{O}_s$ is contained in the strip $\{t \in \mathbb{C} : |\Im(t)| \leq s - 1\}$ in the second to last inequality. We have also the upper bound

$$|a(x,z)| \leq \bar{a}(x) + \Big(\sum_{j\geq 1} |z_j||\psi_j(x)|\Big)^2 \leq \|\bar{a}\|_L + \Big(\sum_{j\geq 1} \rho_j b_j\Big)^2 \leq \|\bar{a}\|_L + 2\Big(\sum_{j\geq 1}(\rho_j - 1)b_j\Big)^2 + 2\Big(\sum_{j\geq 1} b_j\Big)^2 \leq R. \tag{2.5.8}$$

We therefore have

$$0 < r \leq \Re(a(x,z)) \leq |a(x,z)| \leq R < +\infty, \qquad x \in D, \quad z \in \mathcal{O}_\rho. \tag{2.5.9}$$

This uniform ellipticity inequality combined with the holomorphy of $z \mapsto a(z)$ in each variable in $\mathcal{O}_\rho$, implies the holomorphy of the sesquilinear form B and validate the assumption of Definition 2.3.2. The first framework then applies.

In the present setting, the Taylor coefficients of the solution map u can be easily computed using the recursion (2.3.14). Since the diffusion coefficient a is quadratic, then it is easily checked that

$$\partial_{e_i+e_j} a(0) = 2\psi_j\psi_i \quad \text{for} \quad i,j \geq 1, \qquad \text{and} \qquad \partial_\mu a(0) = 0 \quad \text{if} \quad |\mu| \neq 2. \tag{2.5.10}$$

We therefore only retain the indices μ such that $|\nu-\mu|=2$ in (2.3.14). For such indices we have $\nu-\mu=e_i+e_j$ so that $(\nu-\mu)!=1+\delta_{i,j}$. Since $\frac{2}{1+\delta_{i,j}}=(2-\delta_{i,j})$, then according to (2.3.14) the Taylor coefficients can be computed using the recursion

$$\int_D \overline{a}(x)\nabla t_\nu(x)\overline{\nabla w(x)}dx = -\sum_{i,j:e_i+e_j<\nu}(2-\delta_{i,j})\int_D \psi_i(x)\psi_j(x)\nabla t_{\nu-e_i-e_j}(x)\overline{\nabla w(x)}dx, \qquad v\in W. \tag{2.5.11}$$

However, it is unlikely that the Taylor series converges toward the solution map u in the uniform sense. For example, in the simple case where $\overline{a}$ and the ψ_j are constants with $\overline{a}=1$, $\psi_1=5$ and $\psi_j=0$ for any $j\ge 0$, then we have

$$u(y)=\frac{u(0)}{1+25y_1^2}, \qquad y\in U. \tag{2.5.12}$$

It is well known that the map $t\mapsto\frac{1}{1+25t^2}$ is not the sum of its Taylor series in 0.

Log-normal diffusion coefficient:

For the second example, we assume similarly that the sequence $b:=(\|\psi_j\|_L)_{j\ge1}$ belongs to $\ell^p(\mathbb{N})$ for some $p<1$ where $L=L^\infty(D)$. The uniform ellipticity assumption over U is satisfied with

$$r_0:=\exp\Big(-\|b\|_{\ell^1(\mathbb{N})}\Big) \quad\text{and}\quad R_0:=\exp\Big(\|b\|_{\ell^1(\mathbb{N})}\Big). \tag{2.5.13}$$

Now let $0<\varepsilon<\frac{\pi}{2}$ and ρ a sequence with the usual assumption. Given $x\in D$, $z\in\mathcal{O}_\rho$ and $y\in U$ such that $|z_j-y_j|<\rho_j-1$, we have

$$\Re\Big(a(z,x)\Big)=a(y,x)\exp\Big(\sum_{j\ge1}\Re(z_j-y_j)\psi_j(x)\Big)\cos\Big(\sum_{j\ge1}\Im(z_j-y_j)\psi_j(x)\Big)\Big)\ge r_0\exp(-\varepsilon)\cos(\varepsilon), \tag{2.5.14}$$

where we have used $|\sum_{j\ge1}(z_j-y_j)\psi_j(x)|\le\sum_{j\ge1}(\rho_j-1)b_j\le\varepsilon$. By the same argument, we have the upper bound

$$|a(z,x)|=a(y,x)\exp\Big(\sum_{j\ge1}\Re(z_j-y_j)\psi_j\Big)\le R_0\exp(\varepsilon). \tag{2.5.15}$$

We therefore have

$$0<r\le\Re(a(x,z))\le|a(x,z)|\le R<+\infty, \qquad x\in D,\ \ z\in\mathcal{O}_\rho, \tag{2.5.16}$$

with $r=r_0\exp(-\varepsilon)\cos(\varepsilon)$ and $R=R_0\exp(\varepsilon)$. Similar to the first example, Theorem 2.3.3 applies for this second model.

2.5.2 Model (ii): Linear parabolic PDEs with non-affine parametric coefficients

For the parabolic equation (2.1.8) in model (ii), with coefficient a as in (2.5.1), and with the choice of spaces

$$V := L^2(0,T;X)\cap H^1(0,T;X^*) \quad \text{and} \quad W := L^2(0,T;X)\times L^2(D), \quad \text{with} \quad X := H_0^1(D), \tag{2.5.17}$$

the sesquilinear and antilinear forms corresponding to the parabolic problem (2.1.8) read: for $v \in V$ and $w = (w_1, w_2) \in W$

$$B(v,w,z) = \int_0^T \int_D \left(\partial_t v(x,t)\overline{w_1(x,t)} + a(x,z)\nabla_x v(x,t)\overline{\nabla_x w_1(x,t)}\right) dxdt + \int_D v(x,0)\overline{w_2(x)}dx, \tag{2.5.18}$$

and

$$F(w) = \int_0^T \int_D f(x,t)\overline{w_1(x,t)}dxdt + \int_D u_0(x)\overline{w_2(x)}dx, \tag{2.5.19}$$

with all integrals to be understood as the corresponding duality pairings. The boundedness (2.3.4) of these forms is readily verified with the above choices of spaces. The verification of the inf-sup conditions (2.3.5) for the parametric coefficients (1.1.2) or (2.5.1), on the parameter domain $\mathcal{O}_\rho$ follows from the fact that

$$0 < r < \Re(a(x,z)) \le |a(x,z)| \le R, \quad x \in D, \quad z \in \mathcal{O}_\rho, \tag{2.5.20}$$

and using the general arguments given in [73, Appendix].

The application of the previous arguments for the three models studied so far is tied to the simple formula of the diffusion coefficient a and may be tedious when applied to diffusion coefficients with complicated formulas. One can overcome this difficulty by using the second framework. More precisely let us consider a diffusion coefficient a that depends on y according to

$$a(y) = \mathcal{A}(h(y)), \quad h(y) := \sum_{j\ge 1} y_j\psi_j(x), \tag{2.5.21}$$

where $\mathcal{A}$ is a map from $L^\infty(D)$ into itself such that

$$0 < r \le \mathcal{A}(h) \le R < \infty, \qquad h \in h(U), \tag{2.5.22}$$

and such that $\mathcal{A}$ is continuously differentiable over $L^\infty(D)$ viewed as a Banach space over $\mathbb{C}$. We also assume that $(\|\psi_j\|_{L^\infty})_{j\ge 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$. The two examples (2.5.1) correspond to

$$\mathcal{A}(h) = \overline{a} + h^2 \quad \text{and} \quad \mathcal{A}(t) = \exp(h), \quad h \in h(U). \tag{2.5.23}$$

To cast model (i) into the second framework, we introduce the operator

$$\mathcal{P}(u,h) = -\operatorname{div}(\mathcal{A}(h)\nabla u) - f, \tag{2.5.24}$$

This operator is well defined and continuously differentiable from $V \times L$ into W where

$$(V,L,W) := (H_0^1(D), L^\infty(D), H^{-1}(D)), \tag{2.5.25}$$

viewed as complex Banach spaces. For any $u \in V$ and $h \in L$,

$$\frac{\partial \mathcal{P}}{\partial u}(u,h)(v) = -\operatorname{div}(\mathcal{A}(h)\nabla v), \tag{2.5.26}$$

and therefore the uniform ellipticity assumption (2.5.22) implies that $\frac{\partial \mathcal{P}}{\partial u}(u(h(y)),h(y))$ is an isomorphism from V onto W, for all $y \in U$. Therefore, all the assumptions of Theorem 2.4.3 hold. Similar arguments apply for the parabolic problem of model (ii) with

$$\mathcal{P}(u,h) = (\partial_t u - \operatorname{div}(\mathcal{A}(h)\nabla u) - f, u(\cdot,0)), \tag{2.5.27}$$

with the choices

$$V := L^2(0,T;X) \cap H^1(0,T;X^*), \qquad L := L^\infty(D), \qquad W := L^2(0,T;X^*) \times H \tag{2.5.28}$$

where $X := H_0^1(D)$ and $H := L^2(D)$.

2.5.3 Model (iii): Nonlinear elliptic PDE

The nonlinear equation (2.1.11) is associated to the operator,

$$\mathcal{D}(u,y) := u^{2q+1} - \operatorname{div}(a(y)\nabla u) - f, \tag{2.5.29}$$

where $f \in H^{-1}(D)$ with D is a bounded Lipschitz subdomain of $\mathbb{R}^m$. Here $a(y)$ depends affinely on y as in (1.1.2) and satisfies a uniform ellipticity assumption (1.1.3). We assume that $q \geq 0$ is an integer such that $q < \frac{m}{m-2}$ so that $u^{2q+1} \in H^{-1}(D)$ if $u \in H_0^1(D)$. With $V = H_0^1(D)$, we thus have that $\mathcal{D}$ maps $V \times U$ into $V^* = H^{-1}(D)$. More generally, we can consider equations (2.1.1) associated with an operator of the form

$$\mathcal{D}(u,y) := g(u) - \operatorname{div}(\mathcal{A}(h(y))\nabla u) - f, \tag{2.5.30}$$

where $f \in V^*$ and $h(y)$ and $\mathcal{A}$ are as in the previous section. We assume that $(\|\psi_j\|_{L^\infty})_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$. In addition, we assume that g is a function defined on $\mathbb{C}$, such that

1) g is holomorphic on $\mathbb{C}$.

2) $g(0) = 0$ and for any $t \in \mathbb{R}$, $g'(t) \geq 0$.

3) g maps continuously V into V^*.

4) For any $u \in V$, the sesquilinear form $(v, w) \mapsto \int_D g'(u) v \overline{w}$ is continuous over V^2.

These assumptions are in particular fulfilled by the polynomial nonlinearity $g : t \mapsto t^{2q+1}$ when $q < \frac{m}{m-2}$.

Let us now verify the assumptions of Theorem 2.4.3. First, we establish for every $y \in U$, the well-posedness of the nonlinear problem on V understood as a Banach space over $\mathbb{R}$. It follows from the above items 2) and 3) that, for any fixed $y \in U$, the nonlinear operator

$$T(y) : u \mapsto g(u) - \operatorname{div}(\mathcal{A}(h(y))\nabla u), \tag{2.5.31}$$

is continuous, strongly monotone and coercive from V into V^*. By the theory of monotone operators on Banach spaces V over the coefficient field $\mathbb{R}$, see for example Theorem 1 in Chapter 6 of [72], for every $y \in U$ the problem (2.1.1) admits a unique (real-valued) solution $u(y) \in V$.

We next view the spaces (V, L, W) defined as in (2.5.25) as Banach spaces over $\mathbb{C}$ and observe that the map

$$(v, h) \mapsto \mathcal{P}(v, h) := g(v) - \operatorname{div}(\mathcal{A}(h)\nabla v) - f, \tag{2.5.32}$$

is continuously differentiable over $V \times L$, thanks to the assumptions on g and $\mathcal{A}$. For every $(v, h) \in V \times L$, the first partial differential is given by

$$\frac{\partial \mathcal{P}}{\partial u}(v, h)(w) = g'(v)w - \operatorname{div}(\mathcal{A}(h)\nabla w) \in W . \tag{2.5.33}$$

In particular, for any $h \in h(U)$, we have

$$\frac{\partial \mathcal{P}}{\partial u}(u(h), h)(w) = g'(u(h))w - \operatorname{div}(\mathcal{A}(h)\nabla w) . \tag{2.5.34}$$

This operator is associated to the sesquilinear form

$$b(v, w) = \int_D g'(u(h)) v \overline{w} + \int_D \mathcal{A}(h)\nabla v \cdot \overline{\nabla w}. \tag{2.5.35}$$

which is continuous by the upper inequality in (2.5.22) and item 4). In addition it satisfies the coercivity condition

$$b(v, v) \geq r\|v\|_V^2, \quad v \in V, \tag{2.5.36}$$

by the lower inequality in (2.5.22) and item 2). Therefore, by Lax-Milgram theory, it is an isomorphism from V onto W. All the assumptions in Theorem 2.4.3 are thus fullfilled.

Remark 2.5.1

In the case of the nonlinear equation (2.1.11), a possible way to extend the solution for complex valued parameter z would be to rather consider the equation

$$|u|^{2q}u - \operatorname{div}(a(z)\nabla u) = f. \tag{2.5.37}$$

It is easily seen that monotone operator theory applied to the equation verified by the vector (v, w) where $u = v + iw$ allows us to uniquely define the solution $u(z)$ of the above equation under the ellipticity condition $0 < r \leq \Re(a(z)) \leq |a(z)| \leq R$. However the presence of the modulus $|u|$ in the equation obstructs holomorphic dependence on the z_j variable. In our approach, we maintain the original equation (2.1.11). In this case the existence and holomorphy of the solution $u(z)$ for the complex argument z does not *follow from monotone operator theory, but rather from the implicit function theorem argument used in Theorem 4.2.*

2.5.4 Model (iv): Parametrized domain

As a simple example of PDE set on a parametrized domain, we consider the Laplace equation

$$-\Delta v = f \tag{2.5.38}$$

with homogeneous Dirichlet boundary condition set on a physical domain $D(y) \subset \mathbb{R}^2$ that depends on $y \in U$ in the following manner

$$D(y) := \Big\{(x_1, x_2) \ : \ 0 \leq x_1 \leq 1, \ 0 \leq x_2 \leq \phi(x_1, y)\Big\}, \tag{2.5.39}$$

with

$$\phi(t, y) := \overline{\phi}(t) + \sum_{j\geq 1} y_j \psi_j(t), \tag{2.5.40}$$

where the functions $\overline{\phi}$ and ψ_j belong to $W^{1,\infty}([0,1])$, in other words they are Lipschitz continuous on $[0,1]$. We assume that ϕ satisfies a condition of the same type as (1.1.3), namely

$$0 < r \leq \overline{\phi}(t) + \sum_{j\geq 1} y_j \psi_j(t) \leq R < \infty, \quad t \in [0,1], \quad y \in U. \tag{2.5.41}$$

The lower inequality ensures that the boundary of $D(y)$ is not self-intersecting. We also assume that the above series converges in $W^{1,\infty}([0,1])$, uniformly in $y \in U$, that is

$$\delta := \Big\| |\phi'| + \sum_{j\geq 1} |\psi_j'| \Big\|_{L^\infty([0,1])} < \infty \, . \tag{2.5.42}$$

In the above model, the source term f is fixed independently of y and should therefore be defined on the union of all domains $D(y)$ for $y \in U$. For simplicity, we assume that f is defined over

$$\tilde{D} := [0,1] \times [0,R] \tag{2.5.43}$$

and that $f \in L^2(\tilde{D})$. It follows that $f \in L^2(D(y))$ for any $y \in U$ and

$$\|f\|_{L^2(D(y))} \leq \|f\|_{L^2(\tilde{D})}, \qquad y \in U. \tag{2.5.44}$$

Our strategy for treating this model is the following: we use the bijective map

$$\Phi(y): \ x := (x_1, x_2) \mapsto \Phi(x,y) := (x_1, x_2\phi(x_1,y)), \tag{2.5.45}$$

to transport back the solutions $v(y) \in H^1_0(D(y))$ into the reference domain $D := [0,1]^2$ according to

$$u(y) := v(y) \circ \Phi(y), \tag{2.5.46}$$

meaning that $u(x,y) = v(\Phi(x,y),y)$ for all $x \in D$. We then study the linear elliptic PDE satisfied by $u(y)$ on D. This PDE has matricial diffusion coefficient and source term that depend on y. We then show that under certain conditions on the functions ψ_j, one can establish the $\mathbf{HA}(p,\varepsilon)$ for the solution map $y \mapsto u(y)$, using the framework of Theorem 2.3.3.

A change of variables

Having fixed a parameter $y \in U$, we use in what follows the simpler notation u, v and Φ for $u(y)$, $v(y)$ and $\Phi(y)$. The transformation Φ maps the domain D into $D(y)$ and the boundary ∂D into $\partial D(y)$. The function $v \in H^1_0(D(y))$ is the unique solution of the variational problem:

$$\int_{D(y)} \nabla v \cdot \nabla w = \int_{D(y)} fw, \quad w \in H^1_0(D(y)). \tag{2.5.47}$$

The function $u = v \circ \Phi$ is defined on D, and we have

$$\nabla u(x) = (D_\Phi(x))^t \nabla v(\Phi(x)), \tag{2.5.48}$$

where for $x = (x_1, x_2) \in D$,

$$D_\Phi(x) = \begin{bmatrix} 1 & 0 \\ x_2\phi'(x_1,y) & \phi(x_1,y) \end{bmatrix}, \tag{2.5.49}$$

with the derivative in ϕ' is meant with respect to the variable x_1. Since Φ is Lipschitz continuous on D, it follows that $u \in V := H^1_0(D)$. Pulling back the variational formula

(2.5.47) to the reference domain D using the bijective map Φ, one obtains that u is the unique solution to the variational problem

$$\int_D \Big((D_\Phi^{-1})^t \nabla u\Big) \cdot \Big((D_\Phi^{-1})^t \nabla w\Big)\, J_\Phi = \int_D (f \circ \Phi)\, w\, J_\Phi, \qquad w \in V, \tag{2.5.50}$$

where J_Φ is the Jacobian of the transformation Φ which is given by $J_\Phi(x) = \phi(x_1, y)$ for any $x \in D$. We introduce the maps A and g defined on $D \times U$ by

$$A(x,y) := \phi(x_1,y)(D_\Phi^{-1})(D_\Phi^{-1})^t = \begin{bmatrix} \phi(x_1,y) & -x_2\phi'(x_1,y) \\ -x_2\phi'(x_1,y) & \frac{1+(x_2\phi'(x_1,y))^2}{\phi(x_1,y)} \end{bmatrix}, \tag{2.5.51}$$

and

$$g(x,y) := \phi(x_1,y)(f\circ\Phi)(x) = \phi(x_1,y) f\Big(x_1, x_2\phi(x_1,y)\Big), \tag{2.5.52}$$

and the sesquilinear and antilinear forms $B(y)$ and $F(y)$ defined on V by

$$B(y)(w_1,w_2) := \int_D \Big(A(x,y)\nabla w_1(x)\Big)\cdot \overline{\nabla w_2(x)} dx \quad \text{and} \quad F(y)(w) := \int_D g(x,y)\overline{w(x)}dx. \tag{2.5.53}$$

To be consistent with our previous notations, we use the notations $B(w_1, w_2, y)$ instead of $B(y)(w_1, w_2)$ and $F(w, y)$ instead of $F(y)(w)$. From (2.5.50), we deduce that $u(y) \in V$ is the unique solution to the variational problem

$$B(u(y), w, y) = F(w, y), \qquad w \in V. \tag{2.5.54}$$

This is a linear elliptic PDE with parametric matricial diffusion coefficients and parametric source terms. Our next goal is to discuss under which circumstances the assumptions of Theorem 2.3.3 are satisfied for this problem with $V = H_0^1(D)$. We introduce the sequence $b := (b_j)_{j\geq 1}$, with

$$b_j := \|\psi_j\|_{L^\infty([0,1])} + \|\psi_j'\|_{L^\infty([0,1])} \tag{2.5.55}$$

and assume that $b \in \ell^p(\mathbb{N})$ for some $p < 1$. We propose to use this sequence for the verification of the assumptions of Theorem 2.3.3.

Analyticity of the map F

We first study the antilinear form $w \mapsto F(w, y)$. In view of the assumption that $f \in L^2(\tilde{D})$ and the definition (2.5.52) of g, we have a uniform bound of the form

$$|F(w,y)| \leq C\|w\|_V, \quad w \in V,\ y \in U \tag{2.5.56}$$

where

$$C := C_P \sup_{y\in U} \|g(y)\|_{L^2(D)} \leq C_P R \|f\|_{L^2(\tilde{D})}, \tag{2.5.57}$$

with C_P the Poincaré constant for D. More assumptions on f are needed in order to define an holomorphic extension of F in a neighbourhood of U. One sufficient assumption is that the map

$$x_2 \mapsto f(\cdot, x_2), \tag{2.5.58}$$

from $[0, R]$ to $L^2([0, 1])$ is analytic on $[0, R]$. Note that this assumption imposes smooth dependence of f on the second variable. It holds of course if f is analytic in both variables, for example if f is a constant. Since $[0, R]$ is compact, there exists $\varepsilon_1 > 0$ such that the previous map has an holomorphic and uniformly bounded extension on the domain

$$\mathcal{C}_{\varepsilon_1} := \Big\{\xi \in \mathbb{C} : \quad \text{dist}(\xi, [0, R]) < \varepsilon_1\Big\}. \tag{2.5.59}$$

Let now $\rho := (\rho_j)_{j\geq 1}$ a sequence of numbers strictly greater than 1 satisfying $\sum_{j=1}^{\infty}(\rho_j - 1)b_j \leq \varepsilon_1$. We consider the domain $\mathcal{O}_\rho = \otimes_{j\geq 1}\mathcal{O}_{\rho_j}$ where the definition of the open complex domains $\mathcal{O}_s$ is given in (2.4.6). For $z \in \mathcal{O}_\rho$ and $y \in U$ such that $|z_j - y_j| < \rho_j - 1$ for any $j \geq 1$, one has for any $t \in [0, 1]$

$$|\phi(t, z) - \phi(t, y)| = \Big|\sum_{j\geq 1}(z_j - y_j)\psi_j(t)\Big| < \sum_{j\geq 1}(\rho_j - 1)b_j \leq \varepsilon_1. \tag{2.5.60}$$

Since by (2.5.41), $\phi(t, y) \in [0, R]$, then one has $\phi(t, z) \in \mathcal{C}_{\varepsilon_1}$. It follows that the map $y \mapsto g(y)$ defined from U into $L^2(D)$ admits an holomorphic extension $z \mapsto g(z)$ on the domain $\mathcal{O}_\rho$, defined by

$$g(x, z) := \phi(x_1, z) f(x_1, x_2\phi(x_1, z)). \tag{2.5.61}$$

Consequently, the map $y \mapsto F(y)$ from U to V^* admits a uniformly bounded holomorphic extension on the domain $\mathcal{O}_\rho$, defined by

$$F(z)(w) := \int_D g(x, z)\overline{w(x)}dx. \tag{2.5.62}$$

Analyticity of the map B

We now turn to the study of the bilinear form $(w_1, w_2) \mapsto B(y)(w_1, w_2)$. In view of its definition, this amounts to the study of the map $y \mapsto A(y)$ defined in (2.5.51). We propose to show that this map can be extended by holomorphy to complex domains of the form $\mathcal{O}_\rho = \otimes_{j\geq 1}\mathcal{O}_{\rho_j}$, where the domain $\mathcal{O}_s$ for $s > 1$ is defined as in (2.4.6), and that it stays uniformly bounded in the sense of the spectral norm. This establishes the holomorphy and uniform boundedness of the map B.

The entries of the 2×2 symmetric matrix $A(x, y)$ are

$$\phi(x_1, y), \qquad -x_2\phi'(x_1, y) \qquad \text{and} \qquad \frac{1 + (x_2\phi'(x_1, y))^2}{\phi(x_1, y)} \tag{2.5.63}$$

Since $\phi(x_1, y)$ and $\phi'(x_1, y)$ depends linearly on y, then the map $A : y \mapsto A(y)$ can be extended by holomorphy to complex open domains containing U where the quantities $\phi(x_1, z)$, with z replacing y in (2.5.40), never hit 0. Let $0 < \varepsilon \leq \frac{r}{2}$ where r is the lower bound in (2.5.41) and $\rho := (\rho_j)_{j\geq 1}$ (b, ε)-admissible sequence as in (2.2.2). For $z \in \mathcal{O}_\rho$ and $y \in U$ such that $|z_j - y_j| < \rho_j - 1$ for every j, we have by (2.5.60) that $|\phi(t, z) - \phi(t, y)| \leq \varepsilon$ for any $t \in [0, 1]$, therefore

$$\Re\Big(\phi(t, z)\Big) \geq \phi(t, y) - \varepsilon \geq r - \varepsilon \geq \frac{r}{2}, \quad t \in [0, 1]. \tag{2.5.64}$$

Since x_1 varies in $[0, 1]$, then it follows that the map $y \mapsto A(y)$ admits a holomorphic extension on $\mathcal{O}_\rho$ defined by $z \mapsto A(z)$ with

$$A(z)(x) := A(x, z) = \begin{bmatrix} \phi(x_1, z) & -x_2\phi'(x_1, z) \\ \\ -x_2\phi'(x_1, z) & \frac{1+(x_2\phi'(x_1,z))^2}{\phi(x_1,z)} \end{bmatrix}, \quad x \in D. \tag{2.5.65}$$

In addition, for $z \in \mathcal{O}_\rho$, we have for all $x \in D$,

$$\frac{r}{2} \leq |\phi(x_1, z)| = \Big|\phi(x_1, y) + \sum_{j\geq 1}(z_j - y_j)\psi_j(x_1)\Big| \leq R + \varepsilon \tag{2.5.66}$$

and

$$|\phi'(x_1, z)| = \Big|\phi'(x_1, y) + \sum_{j\geq 1}(z_j - y_j)\psi_j'(x_1)\Big| \leq \delta + \varepsilon, \tag{2.5.67}$$

It follows that the spectral norm of $A(z)$ stays uniformly bounded over $\mathcal{O}_\rho$, for example with the supremum value of the entries of $A(z)$ for all $z \in \mathcal{O}_\rho$, which is smaller than

$$R_\varepsilon := \max\Big(R + \varepsilon, R(\delta + \varepsilon), \frac{2}{r}(1 + (R(\delta + \varepsilon))^2\Big), \tag{2.5.68}$$

which only depends on R, r, δ and ε. As a consequence, the map $y \mapsto B(y)$ from U to $\mathcal{B}(V \times V)$, the space of continuous sesquilinear forms over V, admits a uniformly bounded holomorphic extension on $\mathcal{O}_\rho$, defined by

$$B(w_1, w_2, z) := \int_D \Big(A(x, z)\nabla w_1\Big) \cdot \overline{\nabla w_2}, \qquad w_1, w_2 \in V. \tag{2.5.69}$$

Note that the uniform bound is independent of the choice of ρ. Concerning the uniform inf-sup condition, we propose to establish the stronger property that the sesquilinear

forms $B(z)$ are uniformly coercive on the domains $\mathcal{O}_\rho$, up to restricting the range of ε to a smaller interval than $]0, r/2[$.

We introduce now the notation $y := \Re(z)$ and $s := \Im(z)$. Using (2.5.64), (2.5.66) and (2.5.67), we have for any $t \in [0,1]$ and any $z \in \mathcal{O}_\rho$ that

$$\phi(t,y) = \Re(\phi(t,z)) \geq \frac{r}{2} \quad \text{and} \quad |\phi(t,y)| \leq |\phi(t,z)| \leq R + \frac{r}{2} \quad \text{and} \quad |\phi'(t,y)| \leq |\phi'(t,z)| \leq \delta + \frac{r}{2}. \tag{2.5.70}$$

The symmetric real matrices $A(x,y)$ have determinants equal to 1 and, from the above inequalities, their traces are positive and bounded by

$$C_1 := R + \frac{r}{2} + \frac{2}{r}\Big(1 + (\delta + r/2)^2\Big). \tag{2.5.71}$$

Therefore these matrices are positive definite with coercivity constant $\tilde{r} := 1/C_1$. This implies in particular that

$$|B(w,w,y)| \geq \tilde{r}\|w\|_V^2, \quad w \in V,\ y = \Re(z),\ z \in \mathcal{O}_\rho. \tag{2.5.72}$$

To prove the uniform coercivity of the bilinear forms $B(z)$ on $\mathcal{O}_\rho$, it is therefore sufficient to prove that the parametric sesquilinear forms $B(z) - B(y)$ have norms strictly smaller than $\tilde{r}/2$, uniformly on $\mathcal{O}_\rho$. To verify this, we note that the three entries in the symmetric matrices $(A(x,z) - A(x,y))$ are $\phi(x_1,s)$, $-x_2\phi'(x_1,s)$ and

$$\xi(x,z) := \frac{1 + (x_2\phi'(x_1,z))^2}{\phi(x_1,z)} - \frac{1 + (x_2\phi'(x_1,y))^2}{\phi(x_1,y)}. \tag{2.5.73}$$

Since $\mathcal{O}_\rho$ is contained in the tensorized strip $\otimes_{j\geq 1}\{|\Im(z_j)| \leq \rho_j - 1\}$, the condition on ρ readily implies that the two first entries are bounded by ε. Concerning the third entry, we have

$$\xi(x,z) = \Big(1 + (x_2\phi'(x_1,y))^2\Big)\Big(\frac{1}{\phi(x_1,y) + i\phi(x_1,s)} - \frac{1}{\phi(x_1,y)}\Big) + \frac{2x_2^2\phi'(x_1,y)\phi'(x_1,s) - \phi'(x_1,s)^2}{\phi(x_1,z)}. \tag{2.5.74}$$

Therefore, combining the previous inequalities, we obtain

$$|\xi(x,z)| \leq \Big(1 + (\delta + \frac{r}{2})^2\Big)\frac{\varepsilon}{(\frac{r}{2})^2} + \frac{2(R + \frac{r}{2})\varepsilon + \varepsilon^2}{\frac{r}{2}}. \tag{2.5.75}$$

We conclude that the norms of the matrices $(A(t,z) - A(t,y))$ are uniformly bounded by $C_2\varepsilon$ for some constant C_2 depending on R, r and δ. Up to choosing ε small enough, we have $C_2\varepsilon < \frac{\tilde{r}}{2}$, in which case we have for any $w \in V$

$$|B(w,w,z) - B(w,w,y)| \leq \int_D \Big|\Big((A(x,z) - A(x,y))\nabla w\Big) \cdot \overline{\nabla w}\Big| \leq \frac{\tilde{r}}{2}\int_D |\nabla w|^2, \tag{2.5.76}$$

Therefore, with this value of $r > 0$, for any $z \in \mathcal{O}_\rho$ and for any $w \in V$, one has

$$|B(w, w, z)| \geq \frac{\tilde{r}}{2}\|w\|_V^2 \ . \tag{2.5.77}$$

This uniform coercivity implies both inf-sup conditions (2.3.5) with $V = W = H_0^1(D)$.

To complete the verification of the assumptions of Theorem 2.3.3, we only need to reduce the value of ε so that $\varepsilon \leq \varepsilon_1$ where ε_1 was used in the proof of the analyticity of the anti-linear form $F(z)$.

2.6 Conclusion

In this chapter, we have developed a new paradigm which can be used for the treatment of more general parametric PDE than those treated in Chapter 1. This paradigm yields polynomial approximations of the solution map u with provable algebraic rate even in the infinite dimensional setting $d = \infty$. The polynomials considered for approximation are merely the best n-term series associated with Legendre expansions. As seen in §2.5, our approach applies to various classes of parametric PDEs.

In both Chapter 1 and 2, only the approximation of u by Taylor or Legendre series is studied. In practice, the approximation based on such series are rater unconventional since the Taylor and Legendre coefficients are in general out of reach. Moreover, even if such coefficients are known exactly, the exact identification of the n largest terms in a sequence indexed in the lattice $\mathcal{F}$ or $\mathbb{N}^d$ for d large is a difficult task.

In the remainder of this thesis, we propose practical strategies for polynomial approximation of u, the solution map of elliptic models as in Chapter 1 or more general PDEs as models (i)-(ii)-(iii)-(iv) discussed in this chapter. Part II is concerned with intrusive algorithms, which specifically apply to the linear elliptic model with affine parametric dependence of Chapter 1. In part III, we investigate non-intrusive strategies which can be used for more general parametric PDEs.

Part II

Intrusive adaptive algorithms

Chapter 3

An adaptive algorithm for sparse Taylor approximations

Contents

3.1 Introduction

In this chapter, we study a first intrusive method for the approximation of the parametric elliptic model from Chapter 1. The model is given by the equation (1.1.1) where the diffusion coefficient a depends on the parameter y in an affine manner as in (1.1.2) and satisfies the uniform ellipticity assumption $\mathbf{UEA}(r, R)$ given in (1.1.3). We have already studied in Chapter 1 the theoretical approximation of the solution map

$$y \in U \mapsto u(y) \in V, \tag{3.1.1}$$

where $U := [-1, 1]^{\mathbb{N}}$ and $V := H_0^1(D)$, by its truncated Taylor series. We recall in a nutshell the main result in this direction.

We denote by $\mathcal{F}$ the set of finitely supported multi-indices of infinite length, i.e. $\nu := (\nu_j)_{j\geq 1} \in \mathbb{N}^{\mathbb{N}}$ with $\#\{j : \nu_j \neq 0\} < \infty$, and denote by $0_{\mathcal{F}}$ the null multi-index. We introduce the notations

$$\nu! := \prod_{j\geq 1} \nu_j! \quad \text{and} \quad y^\nu := \prod_{j\geq 1} y_j^{\nu_j}, \qquad \nu \in \mathcal{F}, \quad y \in U, \tag{3.1.2}$$

with $0! = 0^0 = 1$. We studied the summability properties in V of partial sums of the formal Taylor series

$$\sum_{\nu\in\mathcal{F}} t_\nu y^\nu, \qquad t_\nu := \frac{\partial^\nu u(0)}{\nu!} \in V, \tag{3.1.3}$$

and their convergence toward the solution map u. The main result concerning such approximations is Theorems 1.5.5, which it is stated as follows.

Theorem 3.1.1

If the sequence $b := (\|\psi_j\|_{L^\infty})_{j\geq 1}$ belongs to $\ell^p(\mathbb{N})$ for some $0 < p < 1$, then

$$u = \sum_{\nu\in\mathcal{F}} t_\nu y^\nu, \tag{3.1.4}$$

in the unconditional uniform sense and the sequence $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$.

The sequence space $\ell^p_m(\mathcal{F})$ contains $\ell^p(\mathcal{F})$ and is defined in (1.5.5). Working under the assumptions of the previous theorem and denoting by $(\Lambda_n^T)_{n\geq 1}$ a sequence of nested sets of indices corresponding each to the n largest $\|t_\nu\|_V$, we have as in (1.2.14)

$$\Big\|u - \sum_{\nu\in\Lambda_n^T} t_\nu y^\nu\Big\|_{\mathcal{V}_\infty} := \sup_{y\in U} \Big\|u - \sum_{\nu\in\Lambda_n^T} t_\nu y^\nu\Big\|_V \leq \sum_{\nu\notin\Lambda_n^T} \|t_\nu\|_V \leq \|(\|t_\nu\|_V)\|_{\ell^p(\mathcal{F})}(n+1)^{-s}, \quad s = \frac{1}{p} - 1. \tag{3.1.5}$$

Using the previous best n-term approximations, we can then approximate simultaneously all the elements from the solution manifold

$$\mathcal{M} := \Big\{u(y)\ ;\ y \in U\Big\}, \tag{3.1.6}$$

at the cost of computing n functions $t_\nu \in V$ with a convergence rate $(n+1)^{-s}$ and a constant that are independent of the number of parameters y_j, here considered infinite ($d = \infty$). This shows that in principle one can overcome the curse of dimensionality in the approximation of u. In computation, however, the sets Λ_k^T are not known to us and in order to find them we would ostensibly have to compute all the coefficients t_ν and sort their V-norms which is infeasible. In order to obtain computable sequences of index sets, we shall not insist on optimality: we shall say that a nested sequence $(\Lambda_k)_{k\geq 0}$ of finite subsets $\Lambda_k \subset \mathcal{F}$ is *near optimal* in the sense of (3.1.5) if it provides the decay

$$\Big\| u - \sum_{\nu\in\Lambda_k} t_\nu y^\nu \Big\|_{\mathcal{V}_\infty} \leq C \|(\|t_\nu\|_V)\|_{\ell^p(\mathcal{F})} (n+1)^{-s}, \qquad n = n(k) = \#(\Lambda_k). \tag{3.1.7}$$

with C a given constant.

The goal in this chapter is to give a concrete algorithm that adaptively builds near optimal sequence $(\Lambda_k)_{k\geq 0}$ and the corresponding Taylor coefficients $(t_\nu)_{\nu\in\Lambda_k}$ at a cost that scales linearly in $\#(\Lambda_k)$. We should point out that similar programs were developed when solving a *single* PDE by either adaptive wavelet methods [30, 31, 49] or by adaptive finite element methods [46, 67, 13, 78]. 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 ones that we are about to present is the use of a *bulk chasing procedure* in order to build the set Λ_{k+1} from the set Λ_k. However, our present setting is significantly different, since the index sets Λ_k are picked from the infinite dimensional lattice $\mathcal{F}$ and the coefficients associated to each $\nu \in \Lambda_k$ are functions in V instead of real numbers.

Based on the summability property $\ell^p_m(\mathcal{F})$ of the Taylor coefficients, we show in §3.2 that our goal in obtaining near optimality (3.1.7) can be reformulated in a more convenient problem of finding near optimal lower sets for a problem of *Taylor residual reduction*. After that, using recursive relations satisfied by Taylor coefficients and the particular structure of the elliptic problem, namely the affine dependence of a in y and the uniform ellipticity assumption, we establish reduction properties on the Taylor residuals which are essential for the design of subsequent algorithms along the line of ideas in [30, 31, 49].

In §3.3, we propose a first adaptive algorithm and prove that the index sets Λ_k generated by the algorithm are near optimal and satisfy (3.1.7). A defect of this algorithm is that the bulk chasing procedure at step k requires at least the computation of d (the parametric dimension) new coefficients t_ν for the indices ν that are in a certain neighbourhood $\mathcal{M}_k$ of Λ_k. The algorithm is then costly for large values of d and impractical in the case of infinite dimension $d = \infty$ that we consider here.

In §3.4, we remedy this defect by introducing a second algorithm which operates at step k the bulk search on a restricted neighbourhood of Λ_k which is of moderate cardinality even in the case $d = \infty$. We prove that this new realistic algorithm generates also index sets that are near optimal and satisfy (3.1.7).

In §3.5, we study 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 does not depend on the number of computed Taylor coefficients.

In §3.6, we propose alternative non-adaptive and adaptive strategies which are computationally much cheaper than the bulk search strategy, yet might benefit from the anisotropic nature of the problem or exhibit the same features of the bulk search. The effectiveness of these algorithms is demonstrated in our numerical examples since some strategies yield the same convergence rate as the bulk search strategy, yet without complete theoretical justification.

Finally, numerical experiments are presented in §3.7, for finite but high dimensional test cases ($y \in [-1,1]^d$ with d up to 255), using finite element for the spatial discretization. We test the adaptive bulk search strategy and compare it with the non-adaptive and adaptive intuitive strategies discusses in §3.6. These experiments confirm that, without any information based on a-priori analysis, the adaptive approach produces near-optimal sets of active indices in terms of convergence rates. 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.

3.2 Taylor Residual formulation

3.2.1 Near optimality with lower sets

We have shown in Chapter 1, formula (1.3.11), that Taylor coefficients satisfy certain recursive relations which can be obtained by differentiating with respect to y the variational formulation satisfied by the instances $u(y)$ of the solution map u,

$$\int_D a(x,y)\nabla u(x,y)\nabla w(x)dx = \int_D f(x)w(x)dx, \qquad v \in V. \tag{3.2.1}$$

Namely $t_{0_\mathcal{F}} = u(0)$ is the unique solution in V of

$$\int_D \overline{a}(x)\nabla t_{0_\mathcal{F}}(x)\nabla w(x)dx = \int_D f(x)w(x)dx, \quad w \in V, \tag{3.2.2}$$

then for multi-indices $\nu \in \mathcal{F} \setminus \{0\}$, the function t_ν is the unique solution in V of

$$\int_D \overline{a}(x)\nabla t_\nu(x)\nabla w(x) = -\sum_{j:\nu_j\neq 0}\int_D \psi_j(x)\nabla t_{\nu-e_j}(x)\nabla w(x), \qquad w \in V. \tag{3.2.3}$$

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 §3.5 and assume for the moment that they can be all solved exactly at a constant cost. In the light of this assumption, it is readily seen that the computation of all the coefficient $\{t_\nu\}_{\nu\in\Lambda}$, associated with a finite set Λ, by the weak formulations (3.2.3) is linear in cost whenever Λ satisfies

$$\nu \in \Lambda \quad \Rightarrow \quad \nu - e_j \in \Lambda \quad \text{for any } j \geq 1 \text{ such that } \nu_j \neq 0. \tag{3.2.4}$$

This category of index sets is called lower sets, and it was already introduced in Chapter 1, Definition 1.5.1. In such case, we remark that the recursion (3.2.3) determines all the Taylor coefficients t_ν for $\nu \in \Lambda$ uniquely. Determining them requires the successive numerical solution of the "nominal" elliptic problems (3.2.2) with $\#(\Lambda)$ many right hand sides. In particular, for computing numerical approximations of the coefficients $(t_\nu)_{\nu\in\Lambda}$, *only a single discretized, parameter-independent "nominal" elliptic problem* (3.2.2) in the domain D must be solved but with $\#(\Lambda)$ many load cases.

As we have already mentioned in Chapter 1, it should be noted that the category of lower sets for approximation with Taylor series was originally introduced in [22] for the practical reason that we just explained. Then the approximation of u using Taylor series truncated to lower index sets was investigated, which we have already done in Chapter 1.

Working under the assumptions of Theorem 3.1.1, the stronger summability $\ell^p_m(\mathcal{F})$ conclusion allows us to localize the best n-term approximations to lower sets while preserving the rate $(n+1)^{-s}$ in (3.1.5). Indeed, denoting by $(\Lambda_n^{T*})_{n\geq 1}$ a lower realisation associated with the monotone envelope of $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$, see Definition 1.5.3, i.e. a sequence of nested lower sets corresponding each to the n largest elements of the monotone envelope of $(\|t_\nu\|_V)_{\nu\in\mathcal{F}}$, we have as explained in (1.5.8)

$$\Big\|u - \sum_{\nu\in\Lambda_n^{T*}} t_\nu y^\nu\Big\|_{\mathcal{V}_\infty} \leq \sum_{\nu\notin\Lambda_n^{T*}} \|t_\nu\|_V \leq \|(\|t_\nu\|_V)\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s}, \qquad s = \frac{1}{p} - 1. \tag{3.2.5}$$

The sequence $(\Lambda_n^{T*})_{n\geq 1}$ is then near optimal in the sense of best n-term approximation (3.1.5). However, as explained earlier with the sequence $(\Lambda_n^T)_{n\geq 1}$ used for (3.1.5), this sequence is also out of reach.

In view of the practical property of lower sets for the computation of Taylor coefficients and the previous theoretical approximation result, we shall only search for computable sparse Taylor series associated with lower sets. In particular, we consider from now on the decay in (3.2.5) as the dacay we aim for near optimality.

In the next section, we propose to simplify further the near optimality analysis by exploiting the special features of the parametric elliptic problem. We will see that the simplified objective meet the settings of the framework of adaptive wavelet methods [30, 31, 49] which allows us later to design adaptive bulk chasing algorithms.

3.2.2 A quadratic Taylor residual

In view of the the recursive relation (3.2.3) satisfied by Taylor coefficients, we find it convenient to work with the average energy norm

$$\|w\|_{\overline{a}} := \Big(\int_D \overline{a}(x)|\nabla w(x)|^2\Big)^{\frac{1}{2}}, \quad w \in V. \tag{3.2.6}$$

Thanks to the uniform ellipticity assumption $\mathbf{UEA}(r,R)$, given in (1.1.3), considered at $y=0$, this norm is equivalent to the V-norm with

$$\sqrt{r}\|w\|_V \leq \|w\|_{\overline{a}} \leq \sqrt{R}\|w\|_V, \qquad w \in V. \tag{3.2.7}$$

It is then equivalent to search for near optimality using the norm $\|\cdot\|_{\overline{a}}$ instead of $\|\cdot\|_V$. We introduce following abbreviated notation

$$c_\nu := \|t_\nu\|_{\overline{a}} = \Big(\int_D \overline{a}|\nabla t_\nu|^2\Big)^{\frac{1}{2}}. \tag{3.2.8}$$

With a slight abuse, we say that c_ν is the energy of the Taylor coefficient t_ν. Now, for $\Lambda \subset \mathcal{F}$ a given set of indices, we introduce the notations

$$e(\Lambda) := \sum_{\nu\in\Lambda} c_\nu^2, \qquad \sigma(\Lambda) := \sum_{\nu\in\mathcal{F}\setminus\Lambda} c_\nu^2. \tag{3.2.9}$$

Although the polynomials y^ν do not form a Fourier basis of $\mathbb{P}_\mathcal{F}$, we may consider that $e(\Lambda)$ and $\sigma(\Lambda)$ measure the energy of the Taylor coefficients on Λ and on its complement $\mathcal{F}\setminus\Lambda$ respectively. In particular, we call $\sigma(\Lambda)$ the quadratic Taylor residual associated with Λ.

In view of the equivalence (3.2.7), under the assumptions of Theorem (3.1.1), the sequence $c := (c_\nu)_{\nu\in\Lambda}$ belongs to $\ell^p_m(\mathcal{F})$. Therefore, the monotone envelope $\mathbf{c} := (\mathbf{c}_\nu)_{\nu\in\mathcal{F}}$ of the sequence c, defined by

$$\mathbf{c}_\nu := \sup_{\mu\geq\nu} |c_\mu|, \quad \nu \in \mathcal{F} \tag{3.2.10}$$

belongs to $\ell^p(\mathcal{F})$. Denoting $(\Lambda_n^{\overline{a}})_{n\geq 1}$ a lower realisation associated with the sequence $\mathbf{c}$, i.e. a sequence of nested lower sets corresponding each to n largest values of $\mathbf{c}_\nu$, we have from Stechkin lemma 1.2.1 applied with p and $q=1$ that

$$\sum_{\nu\notin\Lambda_n^{\overline{a}}} c_\nu \leq \sum_{\nu\notin\Lambda_n^{\overline{a}}} \mathbf{c}_\nu \leq \|\mathbf{c}\|_{\ell^p(\mathcal{F})}(n+1)^{-s} = \|c\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s}, \qquad s = \frac{1}{p} - 1. \tag{3.2.11}$$

Using the equivalence (3.2.7), it is easily checked that if $(\Lambda_n)_{n\geq 1}$ is a sequence of lower sets that is near optimal in the sense of (3.2.11) with a constant C then $(\Lambda_n)_{n\geq 1}$ is near optimal in the sense of (3.2.5) with constant $C\sqrt{\frac{R}{r}}$.

For reasons that we shall explain in the next section, it is more convenient to work with the squares c_ν^2 than with the values c_ν. We explain accordingly how the analysis of the Taylor residuals can be used for the analysis of near optimality in the sense of (3.2.11), hence near optimality in the sense of the the Taylor approximation (3.2.5).

First, using Stechkin lemma 1.2.1 with this time the values p and $q = 2$, we have

$$\sqrt{\sigma(\Lambda_n^{\overline{a}})} \leq \|\mathbf{c}\|_{\ell^p(\mathcal{F})}(n+1)^{-s^*} = \|c\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s^*}, \qquad s^* := \frac{1}{p} - \frac{1}{2}. \tag{3.2.12}$$

The optimal rate s^* associated with the best n-term approximations of the sequence $(\mathbf{c}_\nu)_{\nu\in\mathcal{F}} \in \ell^2(\mathcal{F}) \subset \ell^p(\mathcal{F})$ is better than the rate $s := 1/p - 1$ associated with best n-term approximations in $\ell^1(\mathcal{F})$, both obtained by the same best n-term sets. For index sets that are only near optimal in $\ell^1(\mathcal{F})$, there is no guarantee that they are optimal in $\ell^2(\mathcal{F})$. In contrast, the following useful result states that the inverse is always true.

Lemma 3.2.1

Let $(\Lambda_k)_{k\geq 1}$ be a sequence of (lower) sets that is near optimal in the sense of (3.2.12) with a constant $C \geq 1$, then $(\Lambda_k)_{k\geq 1}$ is also near optimal in the sense of (3.2.11) with constant $(C+1)$.

Proof: Let $n := \#(\Lambda_k) \geq 2$. We have that

$$\mathcal{F} \setminus \Lambda_k \subset \{\mathcal{F} \setminus \Lambda_n^{\overline{a}}\} \cup (\Lambda_n^{\overline{a}} \setminus \Lambda_k),$$

where $\Lambda_n^{\overline{a}}$ is lower and of cardinality n that we used both in (3.2.11) and (3.2.12). Therefore

$$\sum_{\nu\notin\Lambda_k} c_\nu \leq \sum_{\nu\notin\Lambda_n^{\overline{a}}} c_\nu + \sum_{\nu\in\Lambda_n^{\overline{a}}\setminus\Lambda_k} c_\nu \leq \|c\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s} + \sqrt{n}\sqrt{e(\Lambda_n^{\overline{a}} \setminus \Lambda_k)},$$

where we have used (3.2.11) and Cauchy-Schwartz inequality. Since $e(\Lambda_n^{\overline{a}}\setminus\Lambda_k) \leq \sigma(\Lambda_k)$, and Λ_k is near optimal in the sense of (3.2.12) with constant C, then

$$\sum_{\nu\notin\Lambda_k} c_\nu \leq \|c\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s} + \sqrt{n+1}\, C\, \|c\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s^*} \leq (1+C)\|c\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s}.$$

■

It is readily seen that the previous simple argument can only be used with the value $q = 2$ in order to infer near optimality in $\ell^1(\mathcal{F})$ from near optimality in $\ell^q(\mathcal{F})$. This however meets more than enough our needs.

We now summarize the analysis of the present section, we have that if a sequence $(\Lambda_n)_{n\geq 1}$ of nested lower sets is near optimal in the sense of (3.2.12) with constant C,

then it is near optimal in the sense of the benchmark inequality (3.2.5) with constant $(1+C)\sqrt{R/r}$. From this point on, our goal will be then to build computable lower sets that are near optimal the sense of (3.2.12). In the next section, we explain how the arguments of adaptive wavelet approximations can be applied in order to fulfill this purpose.

3.2.3 Recursive estimates and reduction of Taylor residuals

The analysis of the adaptive algorithm is fundamentally based on the recursion (3.2.3). We introduce the following abbreviated notations

$$d_{\nu,j} := \int_D |\psi_j||\nabla t_\nu|^2, \quad \nu \in \mathcal{F}, \quad j \geq 1, \tag{3.2.13}$$

which reflects to some extents contributions to the norm of the load functions in (3.2.3). Our first result is concerned with the comparison of the energy of the Taylor coefficient t_ν solution of (3.2.3) and such quantities.

Lemma 3.2.2

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

$$\sum_{j\geq 1} d_{\nu,j} \leq (1-\gamma)c_\nu^2 \qquad \text{and} \qquad (1+\gamma)c_\nu^2 \leq \sum_{j:\nu_j\neq 0} d_{\nu-e_j,j}. \tag{3.2.14}$$

where $\gamma := \frac{r}{R}$.

Proof: We have seen in Chapter 1, formula (1.4.2) that $\mathbf{UEA}(r,R)$ implies

$$0 < r \leq \overline{a}(x) - \sum_{j\geq 1} |\psi_j(x)|, \qquad x \in D.$$

This is easily obtained by letting every y_j to be equal $-sign(\psi_j(x))$ in (1.1.3) where sign(t) is the sign of the real number t. Using that $r \leq \overline{a}(x) \leq R$ for any $x \in D$, this implies $\sum_{j\geq 1}|\psi_j| \leq (1-\gamma)\overline{a}$. Multiplying by $|\nabla t_\nu|^2$ and integrating over D, we deduce the first inequality in (3.2.14).

As for the second, we take $v = t_\nu$ in (3.2.3) and use the identity $|\alpha\beta| \leq \frac{1}{2}(\alpha^2+\beta^2)$ with the integrands of the right hand side, we obtain

$$c_\nu^2 \leq \sum_{j:\nu_j\neq 0}\int_D \sqrt{|\psi_j|}|\nabla t_{\nu-e_j}|\sqrt{|\psi_j|}|\nabla t_\nu| \leq \frac{1}{2}\sum_{j:\nu_j\neq 0}\int_D |\psi_j|\,|\nabla t_{\nu-e_j}|^2 + \frac{1}{2}\sum_{j:\nu_j\neq 0}\int_D |\psi_j|\,|\nabla t_\nu|^2.$$

Using the first inequality, we deduce the second. ∎

The previous two estimates turn out to be very useful for the estimation of energies outside of lower set. In particular, they imply that the energy outside of any lower set Λ can be controlled by the energy on a certain neighbourhood of Λ. We first define this type of neighbourhood, then state the lemma.

Definition 3.2.3

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

$$\mathcal{M}(\Lambda) := \Big\{\nu \notin \Lambda \ : \ \exists j > 0 \ : \nu - e_j \in \Lambda\Big\}, \tag{3.2.15}$$

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

$$\mathcal{M}(\Lambda) = \mathcal{C}(\Lambda) \setminus \Lambda \quad \text{where} \quad \mathcal{C}(\Lambda) := \Big\{\nu + e_j : \nu \in \Lambda \text{ and } j \geq 1\Big\}. \tag{3.2.16}$$

We have that $\Lambda \cup \mathcal{M}(\Lambda) = \mathcal{C}(\Lambda)$ is a lower set.

The margin $\mathcal{M}(\Lambda)$ is an infinite set even for Λ finite. Indeed, since Λ finite then all its elements are supported in $\{1, \dots, J\}$ for some J, so that in view of the second definition (3.2.16), for any $\nu \in \Lambda$ all the indices $\nu + e_j$ for $j > J$ belong to $\mathcal{M}(\Lambda)$. In the finite dimensional setting $d < \infty$, the margin is a finite set whenever Λ is finite. The reduction property of the energy outside of lower sets is given in the following.

Lemma 3.2.4

Under the uniform ellipticity assumption **UEA**(r, R)*, for any lower set Λ and its margin $\mathcal{M}$, the energies $\sigma(\Lambda)$ and $e(\mathcal{M})$ defined as in (3.2.9) satisfy*

$$\sigma(\Lambda) \leq \frac{1}{1-\delta} e(\mathcal{M}), \quad \text{with} \quad \delta = \frac{1-\gamma}{1+\gamma} < 1. \tag{3.2.17}$$

where $\gamma = \frac{r}{R}$. In particular, if $\mathcal{S}$ is any subset of $\mathcal{M}$ such that $e(\mathcal{S}) \geq \theta e(\mathcal{M})$ with $\theta \in]0, 1[$, then $\Lambda' := \Lambda \cup \mathcal{S}$ satisfies

$$\sigma(\Lambda') \leq \kappa \sigma(\Lambda), \quad \text{with} \quad \kappa := 1 - \theta(1-\delta) < 1 \tag{3.2.18}$$

Proof: From the definition (3.2.15) of the margin, we have that $\nu \notin \Lambda \cup \mathcal{M}$ implies $\nu - e_j \notin \Lambda$ for any j such that $\nu_j \neq 0$. Therefore, The summation of the second inequality in (3.2.14) over all $\nu \notin \Lambda \cup \mathcal{M}$ yields

$$(1+\gamma)\sigma(\Lambda \cup \mathcal{M}) \leq \sum_{\nu \notin \Lambda \cup \mathcal{M}} \sum_{j : \nu_j \neq 0} d_{\nu - e_j, j} \leq \sum_{\mu \notin \Lambda} \sum_{j \geq 1} d_{\mu, j}.$$

This combined with the first inequality in (3.2.14) implies that $(1+\gamma)\sigma(\Lambda \cup \mathcal{M}) \leq$

$(1-\gamma)\sigma(\Lambda)$, which is equivalent to

$$\sigma(\Lambda \cup \mathcal{M}) \leq \delta\sigma(\Lambda).$$

Using that $\sigma(\Lambda\cup\mathcal{M}) = \sigma(\Lambda)-e(\mathcal{M})$, we infer (3.2.17). As for the contraction property, it follows from (3.2.17) using

$$\sigma(\Lambda \cup \mathcal{S}) = \sigma(\Lambda) - e(\mathcal{S}) \leq \sigma(\Lambda) - \theta e(\mathcal{M}) \leq (1-\theta(1-\delta))\sigma(\Lambda).$$

■

The previous lemma show that the margin $\mathcal{M}$ of any lower set Λ, although very negligible in size compared to $\mathcal{F} \setminus \Lambda$, captures a fraction of the total Taylor energy outside of Λ. Therefore, in order to grow Λ and obtain Λ' with energy outside reduced by some factor, it is sufficient to enrich Λ by $\mathcal{M}$ or any subset $\mathcal{S} \subset \mathcal{M}$ that in turn captures a fraction "bulk" of the energy $e(\mathcal{M})$ of the margin. Note however that in order to grow Λ, while preserving the lower set structure of $\Lambda' = \Lambda \cup \mathcal{S}$, one needs to assume a structural condition on $\mathcal{S}$, namely $\mathcal{S}$ lower in $\mathcal{M}(\Lambda)$, which we describe in the following.

We localize the notion of monotone decreasing sequences and lower 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

$$\mu, \nu \in \mathcal{F}_0 \;\text{ and }\; \mu \leq \nu \Rightarrow a_\nu \leq a_\mu. \tag{3.2.19}$$

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

$$\nu \in \mathcal{F}_1,\; \mu \in \mathcal{F}_0 \;\text{ and }\; \mu \leq \nu \Rightarrow \mu \in \mathcal{F}_1. \tag{3.2.20}$$

In the case where $\mathcal{F}_0$ is lower, this is equivalent to saying that $\mathcal{F}_1$ is lower. If (a_ν) is monotone decreasing on $\mathcal{F}_0$, a set $\mathcal{S}_k$ of indices corresponding to the k-largest a_ν in absolute value with $\nu \in \mathcal{F}_0$ is lower in $\mathcal{F}_0$ whenever it is unique. If it is not unique, there exists at least one realization of such set which is lower in $\mathcal{F}_0$. One easy way to obtain a realization is by sorting the indices of $\mathcal{F}_0$ according to

$$\nu \preceq \mu \quad \text{if and only if} \quad |a_\mu| \leq |a_\nu| \quad \text{and} \quad |\nu| \leq |\mu|, \tag{3.2.21}$$

where $|\nu|$ is the ℓ^1 norm of ν, then take for $\mathcal{S}_k$ the k largest indices in the sorted $\mathcal{F}_0$.

Given a lower set Λ with margin $\mathcal{M}$ and $\mathcal{S} \subset \mathcal{M}$, it is easily checked that

$$\Lambda \cup \mathcal{S} \;\text{ is lower} \quad \Longleftrightarrow \quad \mathcal{S} \;\text{ is lower in } \mathcal{M}. \tag{3.2.22}$$

In particular, in Lemma 3.2.4, in order to grow Λ into a lower set $\Lambda' = \Lambda \cup \mathcal{S}$, one has to consider $\mathcal{S}$ lower in $\mathcal{M}$. The general idea of residual reduction techniques, e.g.

[30, 31, 49], adapted to our notations and the problem of near optimality in (3.2.12), consists in optimizing the trade off between the value of θ and size of the enrichment set $\mathcal{S}$, here constrained to be lower in $\mathcal{M}$, that allows to grow Λ into $\Lambda' = \Lambda \cup \mathcal{S}$ with the following implication satisfied

$$(\#(\Lambda)+1)^{s^*}\sqrt{\sigma(\Lambda)} \leq C\|c\|_{\ell^p_m} \quad \Rightarrow \quad (\#(\Lambda')+1)^{s^*}\sqrt{\sigma(\Lambda')} \leq C\|c\|_{\ell^p_m}, \tag{3.2.23}$$

where in our case $s^* = 1/p - 1/2$ and C the wanted near optimality constant. This way, with a careful choice of successive values θ_n and successive enrichment $\mathcal{S}_n$, one obtains a sequence $(\Lambda_n)_{n\geq 1}$ of nested lower set such that

$$(\#(\Lambda_n)+1)^{s^*}\sqrt{\sigma(\Lambda_n)} \leq C\|c\|_{\ell^p_m}, \tag{3.2.24}$$

hence getting the near optimality in the sense of best residual reduction (3.2.12), and consequently near optimality in the sense of (3.1.7). In the following sections, we investigate the design of algorithms along the lines of these ideas.

3.3 A bulk chasing algorithm

In this section, we show how the idea of bulk chase can be used to propose an adaptive algorithm for generating a sequence of sets (Λ_n) that is near optimal in the sense of (3.2.12). The algorithm that we propose is not numerically feasible, however it will guide us in the construction of more practical algorithms in the following sections. We consider the following algorithm:

Algorithm 3.3.1

Let $0 < \theta < 1$. Define $\Lambda_0 := \{0_{\mathcal{F}}\}$, compute $t_{0_{\mathcal{F}}} := u(0)$ and $c_{0_{\mathcal{F}}} := \|t_{0_{\mathcal{F}}}\|_{\bar{a}}$. For $n = 0, 1, \cdots$ do the following:

- *Given that Λ_n has been built and $(t_\nu)_{\nu\in\Lambda_n}$ have been computed, define $\mathcal{M}_n = \mathcal{M}(\Lambda_n)$ and compute t_ν for $\nu \in \mathcal{M}_n$ by the recursion (3.2.3);*
- *Compute c_ν for every ν in $\mathcal{M}_n$ and the monotone envelope of $(c_\nu)_{\nu\in\mathcal{M}_n}$ inside $\mathcal{M}_n$ defined by*

$$\mathbf{c}_\nu(\mathcal{M}_n) := \sup\{c_\mu \ : \ \mu \geq \nu \text{ and } \mu \in \mathcal{M}_n\}; \tag{3.3.1}$$

- *Compute $\mathcal{S}_n$, the smallest lower set in $\mathcal{M}_n$ associated with the largest $\mathbf{c}_\nu(\mathcal{M}_n)$ such that*

$$e(\mathcal{S}_n) \geq \theta e(\mathcal{M}_n); \tag{3.3.2}$$

- *Set $\Lambda_{n+1} = \Lambda_n \cup \mathcal{S}_n$ and go to step $n+1$;*

We have to point out that giving a finite lower set Λ for which we have already computed t_ν for all the indices $\nu \in \Lambda$, we can directly compute certain t_ν for indices ν in $\mathcal{M}$ the margin of Λ using the recurrence (3.2.3). Indeed, if $\mathcal{I}_1(\mathcal{M})$ is the *immediate margin* of Λ, i.e.

$$\mathcal{I}_1(\mathcal{M}) := \Big\{\nu \notin \Lambda : \forall j \geq 0, \;\; \nu_j \neq 0 \Rightarrow \nu - e_j \in \Lambda\Big\} \subset \mathcal{M}, \tag{3.3.3}$$

then we can compute t_ν for all ν in $\mathcal{I}_1(\mathcal{M})$ since we already know every $t_{\nu-e_j}$ that occurs in (3.2.3). We can then repeat this process and compute t_ν for any ν in $\mathcal{I}_2(\mathcal{M})$ where $\mathcal{I}_2(\mathcal{M})$ is the set of indices ν in $\mathcal{M} \setminus \mathcal{I}_1(\mathcal{M})$ such that $\nu - e_j \in \Lambda \cup \mathcal{I}_1(\mathcal{M})$ whenever $\nu_j \geq 1$. Continuing in this way, we can compute all of the $t_\nu \in \mathcal{M}$.

Let us remark that $\mathcal{I}_1(\mathcal{M})$ is of infinite cardinality because it contains the Kronecker indices e_j associated with the infinitely many coordinates y_j which have not been activated yet. The same holds for $\mathcal{I}_2(\mathcal{M})$ since it contains all the indices $e_i + e_j$ with i, j are such that $e_i \in \Lambda$ and $e_j \notin \Lambda$. Indeed $e_i + e_j - e_l$ is equal to either e_i or e_j which in both cases belongs to $\Lambda \cup \mathcal{I}_1(\mathcal{M})$. It can also be proved that every $\mathcal{I}_j(\mathcal{M})$ contains an infinite number of indices ν such that $|\nu| = j$ and are hence of infinite cardinality. However, there exists only a finite number of $\mathcal{I}_j(\mathcal{M})$. Indeed, for all the indices ν in $\mathcal{M}$, we have that $|\nu|$ is bounded by $J := \max_{\mu\in\Lambda} |\mu| + 1$, hence it is easily checked that $I_j(\mathcal{M}) = \emptyset$ for any $j > J$.

In the algorithm above, the enrichment of the set Λ_n is based on sorting the indices $\nu \in \mathcal{M}_n$ by comparing the values $\mathbf{c}_\nu(\mathcal{M}_n)$ and then enriching Λ_n by adding to it the indices ν in $\mathcal{M}_n$ with the largest $\mathbf{c}_\nu(\mathcal{M}_n)$ until the energy of the enriching set $e(\mathcal{S}_n)$ captures a fraction θ of the energy of the margin $e(\mathcal{M}_n)$. The lower structure of the obtained set Λ_{n+1} is crucial in the analysis, one easy way to obtain it consists on reinforcing the sorting as explained in (3.2.21).

We remark that a more natural way to grow Λ_n is by defining Λ_{n+1} to be the smallest monotone set containing Λ_n and contained in $\Lambda_n \cup \mathcal{M}_n$ that captures the bulk energy and satisfies $e(\Lambda_{n+1} \cap \mathcal{M}_n) \geq \theta e(\mathcal{M}_n)$. We shall see that this strategy also yields the desired reduction result given in Theorem 3.3.2 below and consequently the near optimality for Taylor approximations. However, it is not currently known to us how to efficiently implement the search for such minimal lower set in linear time with respect to the cardinality $\#(\Lambda_{n+1} \setminus \Lambda_n)$.

Algorithm 3.3.1 is not satisfactory for several reasons. A first defect is that we can only solve the boundary value problems (3.2.3) approximately, for example using a finite element discretization. We analyze the additional error induced by this discretization in §3.5. Another problem is that in our infinite dimensional setting the margin $\mathcal{M}_n$ has infinite cardinality, therefore there are infinitely many t_ν to be computed which requires in principle solving infinitely many boundary value problems for the corresponding t_ν. Although this problem does not occur in the finite dimensional setting $d < \infty$, it is still reflected by the fact that the size of $\mathcal{M}_n$ is potentially much larger than that of

Λ_n as d gets large and therefore solving the boundary value problems for all $\nu \in \mathcal{M}_n$ becomes the main source of computational complexity. Although the coefficients t_ν are computed once and for all, one should note that at step $n+1$ one computes the new coefficients t_ν for ν in the set $\mathcal{M}(\Lambda_{n+1}) \setminus \mathcal{M}(\Lambda_n)$ which has a cardinality at least large than d. We deal with these computational problems in §3.4.

For now, we remain with the above algorithm and prove its optimality. We first observe that the contraction property (3.2.18) implies $\sigma(\Lambda_{n+1}) \leq \kappa\sigma(\Lambda_n)$ with κ depends on θ as in (3.2.18), therefore

$$\sigma(\Lambda_n) \leq \kappa^n \sigma(\Lambda_0) \leq \kappa^n c_{0_\mathcal{F}}^2 \tag{3.3.4}$$

The Taylor residuals then converge to 0. Thanks to the ℓ^p_m summability of the sequence $(c_\nu)_{\nu\in\mathcal{F}}$, the decay rate of this convergence can be described using the cardinality of the lower sets Λ_n.

Theorem 3.3.2

Under the assumptions of Theorem 3.1.1, the sets Λ_n generated by Algorithm 3.3.1 satisfies

$$\sqrt{\sigma(\Lambda_n)} \leq C_1 \|(c_\nu)\|_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n)+1)^{-s^*}, \quad s^* := \frac{1}{p} - \frac{1}{2}, \tag{3.3.5}$$

where C_1 only depends on (r, R, θ, s^*).

Proof: In order to prove (3.3.5), we first control the cardinality of the updated set $\mathcal{S}_n$. Let μ be the last element added to obtain $\mathcal{S}_n$ in the third step of algortithm 3.3.1 and denote $\mathcal{S} = \mathcal{S}_n - \{\mu\}$. On the one hand, by the optimality of $\mathcal{S}_n$ in the sense of (3.3.2), we have $e(\mathcal{S}) < \theta e(\mathcal{M}_n)$, therefore using the control inequality (3.2.17), we deduce

$$(1-\theta)(1-\delta)\sigma(\Lambda_n) \leq (1-\theta)e(\mathcal{M}_n) < e(\mathcal{M}_n) - e(\mathcal{S}),$$

where δ is given in (3.2.17). One the other hand, $\mathcal{S}$ corresponds to the $(\#(\mathcal{S}_n)-1)$ largest elements of $\mathbf{c}_\nu(\mathcal{M}_n)$. Therefore by Stechkin lemma 1.2.1 with p as in the theorem and $q=2$, we deduce

$$e(\mathcal{M}_n) - e(\mathcal{S}) = \sum_{\nu\in\mathcal{M}_n\setminus\mathcal{S}} c_\nu^2 \leq \sum_{\nu\in\mathcal{M}_n\setminus\mathcal{S}} |\mathbf{c}_\nu(\mathcal{M}_n)|^2 \leq \|(c_\nu)\|^2_{\ell^p_m(\mathcal{F})} (\#(\mathcal{S}_n))^{-2s^*},$$

with s^* is as in the theorem. Combining the two inequalities, we deduce

$$\#(\mathcal{S}_n) \leq C\sigma(\Lambda_n)^{-1/(2s^*)} \quad \text{with} \quad C := [(1-\delta)(1-\theta)]^{-\frac{1}{2s^*}} \|c_\nu\|^{\frac{1}{s^*}}_{\ell^p_m(\mathcal{F})}.$$

Now, from the contraction property (3.2.18), $\sigma(\Lambda_{n+1}) \leq \kappa\sigma(\Lambda_n)$ with $\kappa < 1$ only depending on θ, r and R, hence $\sigma(\Lambda_n) \leq \kappa^{n-k}\sigma(\Lambda_k)$ for any $k \leq n$. Therefore

$$\#(\Lambda_n) \leq \#(\Lambda_0) + \sum_{k=0}^{n-1} \#(\mathcal{S}_k) \leq 1 + C\sum_{k=0}^{n-1} \sigma(\Lambda_k)^{-1/(2s^*)} \leq 1 + C\sigma(\Lambda_n)^{-1/(2s^*)} \sum_{k=0}^{n-1} \kappa^{\frac{n-k}{(2s^*)}} \leq 1 + C'\sigma(\Lambda_n)^{-1/(2s^*)},$$

where $C' := C\frac{\kappa^{\frac{1}{(2s^*)}}}{1-\kappa^{\frac{1}{(2s^*)}}}$. Since $\#(\Lambda_n) \geq 2$, the last inequality implies

$$\sqrt{\sigma(\Lambda_n)} \leq (C')^{s^*}(\#(\Lambda_n)-1)^{-s^*} \leq (C')^{s^*}(2(\#(\Lambda_n)+1))^{-s^*}.$$

We complete the proof by remarking $\left(\frac{C'}{2}\right)^{s^*} = C_1\|c_\nu\|_{\ell^p_m(\mathcal{F})}$ with

$$C_1 := \frac{\sqrt{\kappa}}{2^{s^*}(1-\kappa^{1/(2s^*)})^{s^*}}[(1-\delta)(1-\theta)]^{-1/2}, \tag{3.3.6}$$

Since $\kappa = 1-\theta(1-\delta)$, then the quantity C_1 only depends on r, R, θ and s^*. ■

We have shown that the sets Λ_n generated by the algorithm are near optimal in the sense of (3.2.12). In view of the discussion given at the end of §3.2.2, they also drive near optimal approximations of u in the sense of best n-term approximation by lower sets (3.2.5) with constant $(1+C_1)\sqrt{R/r}$.

Although Algorithm 3.3.1 yields satisfactory convergence results, it can not be implemented in practice. The reason is that the number of the new boundary value problems to solve at each iteration is infinite. In the following section, we investigate strategies for truncating the margin $\mathcal{M}$ of a lower set Λ into a finite restricted margins $\mathcal{N}$ whose energy can still control the energy outside of Λ as in (3.2.17). This shall allow us to modify Algorithm (3.3.1) to a more realistic algorithm and yet obtain the same rate of the above theorem.

3.4 A realistic bulk chasing algorithm

We now want to modify Algorithm 3.3.1 in order to restrict the computation of the t_ν to a finite subset of $\mathcal{M}_n$. In view of the procedure used in the design of the algorithm, the most natural way to truncate the margin $\mathcal{M}$ while preserving the bulk chasing approach, consists in finding a finite set $\mathcal{N}$ that is lower in $\mathcal{M}$ and captures a fraction of $e(\mathcal{M})$ the energy of $\mathcal{M}$, say for example, half of the energy, i.e.

$$e(\mathcal{N}) \geq \frac{e(\mathcal{M})}{2}, \tag{3.4.1}$$

then in view of (3.2.17)

$$\sigma(\Lambda) \leq \frac{2}{1-\delta}e(\mathcal{N}), \tag{3.4.2}$$

yielding an energy control which is crucial for the analysis. From this point on, the design of the bulk chase algorithm is similar to the one in the previous section up to the replacement of the margins by the finite restricted margins. In particular, if in Algorithm (3.3.1), we replace the margins $\mathcal{M}_n$ by restricted finite margins $\mathcal{N}_n$ satisfying

$e(\mathcal{N}_n) \geq \frac{1}{2}e(\mathcal{M}_n)$, then using the proof of Theorem 3.3.2, one easily check that the generated index sets Λ_n satisfy (3.3.5) with a constant C_1' obtained from (3.3.6) with $\frac{1-\delta}{2}$ instead of $1-\delta$. As a consequence, these sets would then yield near optimal Taylor series.

The energy of a given margin $\mathcal{M}$ can only be computed if all the Taylor coefficients $\{t_\nu\}_{\nu\in\mathcal{M}}$ are known. The determination of a restricted margin $\mathcal{N}$ requires then knowing them all which is the primal obstruction we intend to avoid. One possible way to overcome this problem consists in using a priori estimates of the energy $e(\mathcal{M})$. For example, by the same arguments of Lemma 3.2.4, we may obtain

$$e(\mathcal{M}) \leq \frac{1-\gamma}{2\gamma}e(\Lambda), \tag{3.4.3}$$

for any monotone set Λ with margin $\mathcal{M}$. Unfortunately, the previous bound is very pessimistic for indices sets Λ_n that grows with n. Indeed, $e(\mathcal{M}_n)$ which is smaller than $\sigma(\Lambda_n)$ decrease to 0 while $e(\Lambda_n)$ grows. One can also use a priori estimates on Taylor coefficients obtained in Chapter 1 for this purpose. However, this requires the computation of a large number of estimates which in addition may not be very sharp. We propose to construct the restricted sets $\mathcal{N}$ using an incremental strategy.

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

$$\mathcal{N} := \mathrm{SPARSE}(\Lambda, (c_\nu)_{\nu\in\Lambda}, \eta), \tag{3.4.4}$$

is a finite subset of $\mathcal{M}$ which is lower in $\mathcal{M}$ and such that

$$e(\mathcal{M}\setminus\mathcal{N}) \leq \eta. \tag{3.4.5}$$

In view of the ℓ^p summability of the sequence $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1}$, the size of the function ψ_j for j large are negligible. Accordingly, incremental polynomials approximation of u tend to not activate the corresponding variables y_j in the first iterations. Consequently, one natural way to construct a restricted, yet representative, margin of a given lower set is by not advancing in the direction e_j for j large. For a given integer $J \geq 1$, we introduce then the following definition of a restricted margin

$$\mathcal{N}_J(\Lambda) := \mathcal{C}_J(\Lambda)\backslash\Lambda, \qquad \mathcal{C}_J(\Lambda) := \Big\{\nu + e_j : \nu \in \Lambda, j \leq J\Big\}. \tag{3.4.6}$$

In view of the definition (3.2.16) of the margin $\mathcal{M}(\Lambda)$, clearly $\mathcal{N}_J(\Lambda)$ is a subset of $\mathcal{M}(\Lambda)$. Moreover, it is easily checked that $\mathcal{C}_J(\Lambda) := \Lambda \cup \mathcal{N}_J(\Lambda)$ is lower, hence in view of the equivalence (3.2.22) the set $\mathcal{N}_J(\Lambda)$ is lower in $\mathcal{M}(\Lambda)$. The set is of finite cardinality with $\#(\mathcal{N}_J) \leq J\#(\Lambda)$. Finally, we have

Lemma 3.4.1

Let Λ be a lower set and $\mathcal{M}$ its margin. If $J \geq 0$ is such that $\sum_{j>J} \|\psi_j\|_{L^\infty} \leq 2\gamma r \frac{\eta}{e(\Lambda)}$ then with the above definition of $\mathcal{N}_J$, one has

$$e(\mathcal{M} \setminus \mathcal{N}_J) \leq \eta. \tag{3.4.7}$$

Proof: We proceed in a similar way to the proof of Lemma 3.2.4. First, using (3.2.14) we write

$$(1+\gamma)e(\mathcal{M} \setminus \mathcal{N}_J) \leq \sum_{\nu \in \mathcal{M}\setminus\mathcal{N}_J} \Big(\sum_{j:\nu_j \neq 0} d_{\nu-e_j,j} \Big).$$

From the definition (3.2.16) of the margin $\mathcal{M}$ and the definition (3.4.6) of the reduced margin $\mathcal{N}_J$, we have

$$\mathcal{M}\setminus\mathcal{N}_J = \Big\{ \mu + e_k : \mu \in \Lambda, k > J \Big\} \setminus \Lambda. \tag{3.4.8}$$

Let $\nu \in \mathcal{M}\setminus\mathcal{N}_J$ that we write $\nu = \mu + e_k$ with $\mu \in \Lambda$ and $k > J$. For $j \neq k$ such that $\nu_j \neq 0$, we have $\nu - e_j = (\mu - e_j) + e_k$ is a sum of $(\mu - e_j) \in \Lambda$, because Λ is lower, and e_k with $k > J$, therefore $\nu - e_j$ belongs necessarily to $\Lambda \cup \{\mathcal{M}\setminus\mathcal{N}_J\}$. If $j = k$, then $\nu - e_j = \mu \in \Lambda$. We can then divide the sum in the the above inequality as

$$(1+\gamma)e(\mathcal{M} \setminus \mathcal{N}_J) \leq \sum_{\nu \in \mathcal{M}\setminus\mathcal{N}_J} \Big(\sum_{\substack{j:\nu_j \neq 0, \\ \nu-e_j \in \Lambda}} d_{\nu-e_j,j} \Big) + \sum_{\nu \in \mathcal{M}\setminus\mathcal{N}_J} \Big(\sum_{\substack{j:\nu_j \neq 0, \\ \nu-e_j \in \mathcal{M}\setminus\mathcal{N}_J}} d_{\nu-e_j,j} \Big).$$

Next we remark that $\nu \in \mathcal{M}\setminus\mathcal{N}_J$ and $\nu - e_j \in \Lambda$ implies necessarily $j > J$, because otherwise we would have $\nu = (\nu - e_j) + e_j$ belongs to $\mathcal{N}_J$ which contradicts $\nu \in \mathcal{M}\setminus\mathcal{N}_J$. In view of this remark and letting $\mu = \nu - e_j$, the previous inequality implies

$$(1+\gamma)e(\mathcal{M} \setminus \mathcal{N}) \leq \sum_{\mu \in \Lambda} \sum_{j>J} d_{\mu,j} + \sum_{\mu \in \mathcal{M}\setminus\mathcal{N}_J} \sum_{j \geq 1} d_{\mu,j}.$$

Using (3.2.14), the last term in the right side is smaller than $(1-\gamma)e(\mathcal{M}\setminus\mathcal{N}_J)$. As for the first, from the definition (3.2.13) of the quantities $d_{\mu,j}$, we infer

$$\sum_{\mu \in \Lambda} \sum_{j>J} d_{\mu,j} = \sum_{\mu \in \Lambda} \int_D \Big(\sum_{j>J} |\psi_j| \Big) |\nabla t_\mu|^2 \leq \Big\| \sum_{j>J} |\frac{\psi_j}{\overline{a}}| \Big\|_{L^\infty} e(\Lambda).$$

We deduce then that

$$2\gamma\, e(\mathcal{M}\setminus\mathcal{N}_J) \leq e(\Lambda) \Big\| \sum_{j>J} |\frac{\psi_j}{\overline{a}}| \Big\|_{L^\infty} \leq \frac{e(\Lambda)}{r} \sum_{j>J} \|\psi_j\|_{L^\infty}, \tag{3.4.9}$$

where we have used the uniform ellipticity assumption at $y = 0$ to get $0 < r \leq \overline{a}$. The proof is complete. ∎

The previous lemma does not implies that $\mathcal{N}_J$ captures directly a fraction of the energy $e(\mathcal{M})$. We propose an incremental strategy

$$(\mathcal{N}, \eta) := \text{OVERGROW}(\mathcal{M}, (c_\nu)_{\nu\in\mathcal{M}}, \theta), \tag{3.4.10}$$

which giving $\mathcal{M}$ the margin of Λ, output the value η and a restricted margin $\mathcal{N}$ such that $e(\mathcal{M}\setminus\mathcal{N}) \leq \eta$ and captures at least a fraction θ of the energy $e(\mathcal{M})$. For example, using the restricted margin defined in (3.4.6), this can be done by incrementing J, and accordingly growing $\mathcal{N}_J$ until we captures the desired fraction.

Algorithm 3.4.2

Let Λ be a lower set, $\mathcal{M}$ its margin, $\theta \in]0,1[$ and $\eta > 0$. Let $j = 0$, then do the following:

- *Define $\eta_j := 2^{-j}\eta$ and $\mathcal{M}_j := \text{SPARSE}(\Lambda, (c_\nu)_{\nu\in\Lambda}, \eta_j)$;*
- *Compute t_ν and c_ν for $\nu \in \mathcal{M}_j$ and compute $e(\mathcal{M}_j)$;*
- *If $e(\mathcal{M}_j) < \frac{2(2-\theta)}{1-\theta}\eta_j$, then go directly to step $j+1$;*
- *Else, terminate the loop in j, and output the set $\mathcal{M}_j$ and the value η_j.*

We have

$$e(\mathcal{M}_j) \geq \theta e(\mathcal{M}) \tag{3.4.11}$$

Proof: The previous loop always terminates. Indeed, η_j decrease to 0, while the energies of the restricted margins $e(\mathcal{M}_j)$ increase. Let J the last integer in the previous loop. One the one hand $e(\mathcal{M}_J) \geq \frac{2(2-\theta)}{1-\theta}\eta_J$, therefore

$$e(\mathcal{M}_J) \geq \theta e(\mathcal{M}_J) + 2(2-\theta)\eta_J \geq \theta e(\mathcal{M}_J) + (2-\theta)\eta_J.$$

One the other hand $e(\mathcal{M}\setminus\mathcal{M}_J) \leq \eta_J$, it follows that $e(\mathcal{M}_J) \geq e(\mathcal{M}) - \eta_J$, hence

$$e(\mathcal{M}_J) \geq \theta(e(\mathcal{M}) - \eta_J) + (2-\theta)\eta_J \geq \theta e(\mathcal{M}) + 2(1-\theta)\eta_J \geq \theta e(\mathcal{M}),$$

which finishes the proof. ■

We now consider the following algorithm:

Algorithm 3.4.3

Let $0 < \theta < 1$. Define $\Lambda_0 := \{0_\mathcal{F}\}$, compute $t_{0_\mathcal{F}} := u(0)$, $c_{0_\mathcal{F}} = \|t_{0_\mathcal{F}}\|_{\overline{a}}$ and set $\eta_0 = c_{0_\mathcal{F}}$. For the values $n = 0, 1, \cdots$, do the following

- *Given that Λ_n has been defined and $(t_\nu)_{\nu\in\Lambda_n}$ have been computed, define $\mathcal{M}_n = \mathcal{M}(\Lambda_n)$.*

- *Output the reduced margin* $(\mathcal{M}_{j_n}, \eta_{j_n}) := \mathrm{OVERGROW}(\mathcal{M}_n, (c_\nu)_{\nu\in\mathcal{M}_n}, \theta)$ *and define* $\eta_{n+1} := \eta_{j_n}$.
- *Define*
$$\mathbf{c}_\nu(\mathcal{M}_{j_n}) := \sup\{c_\mu \ : \ \mu \geq \nu \text{ and } \mu \in \mathcal{M}_{j_n}\},$$
and compute $\mathcal{S}_n$, *the smallest lower set in* $\mathcal{M}_{j_n}$ *associated with the largest* $\mathbf{c}_\nu(\mathcal{M}_{j_n})$ *such that*
$$e(\mathcal{S}_n) \geq \frac{1}{2}\, e(\mathcal{M}_{j_n});$$
- *Set* $\Lambda_{n+1} = \Lambda_n \cup \mathcal{S}_n$ *and go to step* $n+1$;

At every step of the algorithm, we have $e(\mathcal{M}_{j_n}) \geq \theta e(\mathcal{M}_n)$, therefore in view of (3.2.17)

$$\sigma(\Lambda_n) \leq \frac{1}{\theta(1-\delta)} e(\mathcal{M}_{j_n}). \tag{3.4.12}$$

Then considering the restricted margin $\mathcal{M}_{j_n}$ instead of $\mathcal{M}_n$ preserve the key features of the bulk chasing algorithm. By exactly the same proof of Theorem 3.3.2, yet with $\theta(1-\delta)$ instead of $(1-\delta)$ and $\frac{1}{2}$ instead of θ, we can prove that the previous algorithm yields satisfactory results. We have

Theorem 3.4.4

Under the assumptions of Theorem 3.1.1, the sets Λ_n *generated by the previous algorithm satisfy*

$$\sqrt{\sigma(\Lambda_n)} \leq C_2 \|(c_\nu)\|_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n)+1)^{-s^*}, \quad s^* := \frac{1}{p} - \frac{1}{2}, \tag{3.4.13}$$

where C_2 *only depends on* (r, R, θ, t).

Although Algorithm 3.4.3 meets the benchmark of the optimal rate (3.2.12) under the minimal assumptions of Theorem 3.1.1, 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 (3.2.12) 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 $\sigma(\Lambda)$, then the cardinality of the produced

set $\mathcal{N}$ should be bounded by the optimal rate

$$\#(\mathcal{N}) \leq C\|(c_\nu)\|^{1/s^*}_{\ell^p_m(\mathcal{F})}\eta^{-1/(2s^*)}. \tag{3.4.14}$$

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 $\eta^{-1/(2s^*)}$, and on the other hand the number J that ensures (3.4.7) is of the order $\eta^{-1/s}$ where $s = \frac{1}{p} - 1 = s^* - \frac{1}{2}$. Therefore $\eta^{-1/(2s^*)}$ in (3.4.14) is a-priori replaced by the non-optimal rate $\eta^{-1/(2(s^*)^2-s^*)}$

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 obtained in Chapter 1 in order to control the energy outside of the set $\mathcal{N}$. 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 §3.6.

3.5 Space discretization

In practice, we set a target accuracy $\varepsilon > 0$ and design the bulk chase procedures in such a way that the algorithm terminates when $\sigma(\Lambda_n) \leq \varepsilon$. In this section, we analyse the additional error which occurs due to finite element spatial discretization, and therefore a relevant choice for ε is an estimated value of this additional error, such as given by standard residual-based finite element error estimator.

The boundary value problems that recursively give the Taylor coefficients t_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* t_ν and still retain the performance of Algorithm 3.3.1 and Algorithm 3.4.3.

For the purpose of simplicity, we consider here the situation where the same spatial discretization is used for all t_ν. However, the analysis in §8 of [34] 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 map $y \mapsto u_h(y) \in V_h$, with each $u_h(y)$ solution to

$$\int_D a(x,y)\nabla u_h(y)\nabla w_h = \int_D f v_h \qquad \forall w_h \in V_h \ . \tag{3.5.1}$$

By assumption **UEA**(r,R), for any closed subspace $V_h \subset V$ the Finite Element approximation is uniquely defined and the analysis in [34, 33], which is recalled also in Chapter 1, and all results of the present analysis 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.

$$u_h(y) = \sum_{\nu\in\mathcal{F}} t_{\nu,h}\, y^\nu, \qquad \text{where} \qquad t_{\nu,h} := \frac{\partial_\nu u_h(0)}{\nu!} \in V_h. \tag{3.5.2}$$

Moreover, similarly to the norms $\|t_\nu\|_V$, the norms $\|t_{\nu,h}\|_V$ can be estimated by the same bound (1.4.19) which use only the uniform ellipticity assumptions

$$\|t_{\nu,h}\|_V \le \inf_{0<\delta<r} \left\{ \frac{\|f\|_{V_h^*}}{\delta} \inf\{\rho^{-\nu} : \rho \text{ is } \delta\text{-admissible}\} \right\} \tag{3.5.3}$$

This consequently leads to a result similar to Theorem 3.1.1.

Theorem 3.5.1

Under the assumptions of Theorem 3.1.1, the sequence $(\|t_{\nu,h}\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$. Moreover the norm $\|(\|t_{\nu,h}\|_V)\|_{\ell^p_m(\mathcal{F})}$ is bounded independently of h.

The coefficients $t_{\nu,h}$ can be computed recursively by solving linear systems corresponding to the space-discretized boundary value problems. Indeed, by differentiating the variational formula (3.5.1) at $y = 0$, we obtain that the descretized Taylor coefficients $t_{\nu,h} \in V_h$ are the solution to the elliptic boundary value problems given in weak form by: $t_{0_\mathcal{F},h} := u_h(0)$ satisfies

$$\int_D \overline{a}(x)\nabla t_{0_\mathcal{F},h}(x)\nabla w_h(x)dx = \int_D f(x)w_h(x)dx, \quad w_h \in V_h, \tag{3.5.4}$$

and the others coefficients satisfy the recursions

$$\int_D \overline{a}(x)\nabla t_{\nu,h}(x)\nabla w_h(x)dx = -\sum_{j:\nu_j\neq 0} \int_D \psi_j(x)\nabla t_{\nu-e_j,h}(x)\nabla w_h(x)dx, \qquad w_h \in V_h. \tag{3.5.5}$$

For the approximate Taylor coefficients, we introduce once more their energies as $c_{\nu,h} := \|t_{\nu,h}\|_{\overline{a}}$. We may define energies $e_h(\Lambda)$ and $\sigma_h(\Lambda)$ associated with discretized Taylor coefficients as in (3.2.9). We introduce the quantity

$$c_{\nu,h} := \|t_{\nu,h}\|_{\overline{a}} = \Big(\int_D \overline{a}|\nabla t_{\nu,h}|^2 \Big)^{\frac{1}{2}}. \tag{3.5.6}$$

We introduce also the quantities $d_{\nu,h,j}$ as in (3.2.13). The lemmas 3.2.2 and 3.2.4 can be proved for the discretized case exactly by the same arguments. Finally, algorithms 3.3.1 or 3.4.3 are given in a similar way, by simply replacing t_ν and c_ν by $t_{\nu,h}$ and $c_{\nu,h}$ respectively. For these algorithms, we obtain the convergence results by the exact same approach as without space discretization.

Theorem 3.5.2

Under the assumptions of Theorem 3.1.1, the application of each of the algorithms 3.3.1 or 3.4.3 in the space discretized setting yields a sequence of sets (Λ_n) that satisfies

$$\sqrt{\sigma_h(\Lambda_n)} \leq C_i \|(c_{\nu,h})\|_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n)+1)^{-s^*}, \quad s^* := \frac{1}{p} - \frac{1}{2}, \tag{3.5.7}$$

where $C_i = C_1$ or C_2 are as in the continuous setting (depending on r, R, θ and on s^*, but being independent of h). Consequently, we have in both cases

$$\Big\| u_h - \sum_{\nu\in\Lambda_n} t_{\nu,h} y^\nu \Big\|_{\mathcal{V}_\infty} \leq \sqrt{\frac{R}{r}} (1+C_i) \|(\|t_{\nu,h}\|_V)\|_{\ell^p_m} (\#(\Lambda_n)+1)^{-s}, \quad s = \frac{1}{p} - 1, \tag{3.5.8}$$

with $i = 1$ or 2.

The previous rate is near optimal in the sense of the benchmark rate (3.2.5). We need only to study the quantity $\|u - u_h\|_{\mathcal{V}_\infty}$. In particular, through the quantification of the space discretization error. The well-known theory of finite elements tells us that the rate of convergence of

$$\|u(y) - u_h(y)\|_V, \tag{3.5.9}$$

in terms of the decay of h is controlled by the smoothness of $u(y)$ 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 for every $y \in U$

$$\|u(y) - u_h(h)\|_V \leq C h^r \|u(y)\|_{H^{1+r}(D)}, \tag{3.5.10}$$

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

Corollary 3.5.3

Under the assumptions of Theorem 3.1.1 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 3.3.1 or 3.4.3 in the space discretized setting, we obtain

$$\Big\| u(y) - \sum_{\nu\in\Lambda_n} t_{\nu,h} y^\nu \Big\|_{\mathcal{V}_\infty} \leq C h^r \sup_{y\in U} \|u(y)\|_{H^{1+r}(D)} + \sqrt{\frac{R}{r}} (1+C_i) \|(c_{\nu,h})\|_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n)+1)^{-s}. \tag{3.5.11}$$

The largest value of r for which $\sup_{y\in U} \|u(y)\|_{H^{1+r}} < \infty$ depends on many consideration: the smoothness of the right hand side f, the smoothness of the diffusion coefficient a and the smoothness of the boundary of the spacial domain D.

As an example, we assume that $f \in L^2(D)$ and the diffusion coefficients $a(y)$ satisfy

$$\sup_{y\in U} \|a(\cdot, y)\|_{W^{1,\infty}(D)} < \infty. \tag{3.5.12}$$

In view of the elliptic model (1.1.1), we have that the functions $u(y)$ satisfy the Poisson equation

$$-\Delta u(y) = \frac{1}{a}\Big[f - \nabla a \cdot \nabla u(y)\Big] \quad \text{in} \quad D\,, \qquad u(y)|_{\partial D} = 0\,. \tag{3.5.13}$$

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

$$W = \big\{v \in V : \Delta v \in L^2(D)\big\}\ . \tag{3.5.14}$$

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 $\frac{1}{2} \le r \le 1$. We refer to [54] for a general treatment of elliptic problems on non-smooth domains.

In the numerical experiment section, we 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 (3.5.12), 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 $0 < r < \frac{1}{2}$ that depends on the maximal contrast R/r, see for example [11].

3.6 Alternative algorithms ($d < \infty$)

Although the algorithms 3.3.1 and 3.4.3 can be implemented in practice, with the first in the finite dimension, we have seen that both can be computationally expansive if the margins or the restricted margins have considerable sizes. Since a given Taylor coefficient is computed once and for all, then at every step n of these algorithms, the number of the newly computed Taylor coefficients is

$$\#\Big(\mathcal{M}_r(\Lambda_n) \setminus \mathcal{M}_r(\Lambda_{n-1})\Big) \tag{3.6.1}$$

where $\mathcal{M}_r$ stand for the margin or restricted margin depending on the algorithm. We recall that $\Lambda_n = \Lambda_{n-1} \cup \mathcal{S}_{n-1}$. In view of the definition of the reduced margin (3.4.6), it is readily seen that given Λ lower and $\mathcal{S}$ lower in $\mathcal{M}(\Lambda)$, that

$$\mathcal{M}_r(\Lambda \cup \mathcal{S}) \setminus \mathcal{M}_r(\Lambda) := \Big\{\nu + e_j : \nu \in \mathcal{S}, j \le J\Big\} \setminus \mathcal{S} \tag{3.6.2}$$

It is obvious that the cardinality of this set is smaller than $J\#(\mathcal{S})$. It can be however very large. It is then crucial to consider algorithms that allow the construction of lower index sets Λ_n in moderate time in $\#(\Lambda_n)$. We propose some algorithms for the generation of the sets Λ_n of "active" Taylor coefficients. We consider many non-adaptive strategies that are based on a-priori choices of the sets Λ_n, and two adaptive strategies that exploit the results of earlier computations. For the sake of notational simplicity, we describe these algorithms without the additional finite element discretization error, therefore using the notation t_ν and c_ν. The adaptation of these stratetgies and algorithms 3.3.1 and 3.4.3 to the finite element setting is straightforward and can be examined as in the previous section.

In view of the previous works on the polynomial approximation of the elliptic model, as discussed in the general introduction, it is of interest to first compare the approximation based on Talyor series with other types of approximation, such as Neumann series, Galerkin projection, interpolation...using as lower index sets those used for each method. First, isotropic sets, namely the isotropic rectangular block, simplex and hyperbolic cross, i.e.

$$\mathcal{B}_k := \Big\{\nu \in \mathcal{F} : \nu_j \leq k\Big\}, \quad \mathcal{S}_k := \Big\{\nu \in \mathcal{F} : \sum_{j=1}^{d} \nu_j \leq k\Big\}, \quad \mathcal{H}_k := \Big\{\nu \in \mathcal{F} : \prod_{j\geq 1}(\nu_j+1) \leq k\Big\}. \tag{3.6.3}$$

These sets are lower but of infinite cardinality for $d = \infty$. They should only be considered in the first J coordinates for J a given finite integer or in the case $d < \infty$ in which $\mathcal{F} = \mathbb{N}^d$. In such case, we recall that the associated multi-variate polynomial spaces $\mathbb{P}_{\mathcal{S}_k}$ and $\mathbb{P}_{\mathcal{B}_k}$ are respectively the space of polynomials of total degree at most k and of polynomials of degree at most k in each variable. The dimensions of the polynomials spaces are

$$\#(\mathcal{H}_k) \simeq k(\log k)^{d-1} \leq \#(\mathcal{S}_k) = \binom{k+d}{k} \leq \#(\mathcal{B}_k) = (k+1)^d. \tag{3.6.4}$$

We observe that these dimensions grow exponentially with the dimension d of y, reflecting the curse of dimensionality. When used, even in a finite setting with $d >> 1$, they should be considered only with few direction $e_1, \ldots, e_J$.

Anisotropic versions of the previous sets can be used in order to take into account the anisotropy of the problem. adopting the notation in [7, 8, 69], yet with a slight normalization difference, we let $\alpha = (\alpha_j)_{j\geq 1}$ be a sequence of strictly positive numbers and then introduce the notations

$$\mathcal{B}_{k,\alpha} := \Big\{\nu : \alpha_j\nu_j \leq k\Big\}, \quad \mathcal{S}_{k,\alpha} := \Big\{\nu : \sum_{j=1}^{d} \alpha_j\nu_j \leq k\Big\}, \quad \mathcal{H}_{k,\alpha} := \Big\{\nu : \prod_{j\geq 1}(\nu_j+1)^{\alpha_j} \leq k\Big\}. \tag{3.6.5}$$

The first set $\mathcal{B}_{k,\alpha}$ define also a rectangular block $\mathcal{B}_\mu := \{\nu : \nu \leq \mu\}$ where $\mu \in \mathcal{F}$ is defined by $\mu_j = \lfloor \frac{k}{\alpha_j} \rfloor$ for every j. We observe that $\alpha = (1, 1, \ldots)$ yields to the isotropic setting.

In contrast to the isotropic sets, the anisotropic versions can be of finite cardinality even in the case $d = \infty$. For instance, taking the values of α_j relatively large for the directions $j > J$, in the sense $\alpha_j > k$ for the block and simplex or $2^{\alpha_j} > k$ for the cross, yields that the directions $J+1, J+2, \ldots$ are not activated in the index sets in which case the latter are if considered in dimension $d = J$. Accordingly, the variables $y_{J+1}, y_{J+2}, \ldots$ are inactive in the polynomial approximation. However, having $(\alpha_j)_{j\geq 1}$ fixed and increasing the value of k, new directions are unlocked, with the new index set at least doubling in cardinality, which also reflects the curse of dimensionality.

The parameter α should reflect the anisotropy of the problem: the smaller is the dependance on the variable y_j, the larger is the value of α_j. For example, for the elliptic model studied in this chapter, an intuitive choice can be given by

$$\alpha_j := \|\psi_j\|^{-1}_{L^\infty(D)}, \qquad j \geq 1. \tag{3.6.6}$$

Since $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p_m(\mathcal{F})$, then $\|\psi_j\|_{L^\infty(D)} \to_{j\to\infty} 0$, so that approximations using the anisotropic sets do not activate, for small values of k, the variables y_j for j large.

The index sets described previously are all lower and known in advance. The computation of the corresponding Taylor series is then linear in $\#(\Lambda)$. However, even in the anisotropic case with α as above, the convergence may not be satisfactory. For instance the previous choice of α might not capture coupling phenomena between the variables y_j. In the following, we propose algorithms that perform extra processing work, but that hopefully yield to an acceleration of the convergence.

3.6.1 Largest estimates algorithm

The norms $\|t_\nu\|_V$ of Taylor coefficients were estimated in Chapter 1 by the bound (1.4.19), which is given by

$$\|t_\nu\|_V \leq h_\nu := \inf_{0<\delta<r} \left\{ \frac{\|f\|_{V^*}}{\delta} \inf\{\rho^{-\nu} : \rho \text{ is } \delta\text{-admissible}\} \right\}; \tag{3.6.7}$$

with δ-admissibility is defined in (1.4.11). It can be then of interest to consider best n-term sets associated with $(h_\nu)_{\nu\in\mathcal{F}}$ for the truncation of Taylor series. Moreover, since the sequence $(h_\nu)_{\nu\in\mathcal{F}}$ is monotone decreasing, it is possible to choose best n-term sets which are lower. The corresponding series will then yield approximations that are near optimal in the sense of (3.2.5). For a discussion, we refer to formula (1.4.22).

The sequence of estimates is in general not easily computable, since it is obtained by a double optimization problem. One simpler estimate can be obtained by eliminating

the optimization on δ, for example taking simply $\delta = \frac{r}{2}$. Using the same notation h_ν, the new estimates are giving by

$$h_\nu := \frac{2\|f\|_{V^*}}{r} \inf\Big\{\rho^{-\nu} : \rho \text{ is } \{\frac{r}{2}\}\text{-admissible}\Big\}, \tag{3.6.8}$$

The new sequence $(h_\nu)_{\nu\in\mathcal{F}}$ is also monotone decreasing since the ρ_j are by definition greater that 1, see (1.4.11). Also, the inspection of the proof of Theorem 1.4.5 shows that $(h_\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p(\mathcal{F})$. The sequence can then be used for best n-term approximation as explained above. However, this sequence can be difficult to compute since it is also obtained by optimization. We distinguish a particular case where the computation is immediate then discuss the general case.

Piecewise constant diffusion coefficients

We consider a simple setting where $\overline{a}$ is constant equal to 1 and the functions ψ_j have non-overlapping supports and are constant and have values $b_j > 0$. The uniform ellipticity assumption $\mathbf{UEA}(r, R)$ is here equivalent to $\max_{j\geq 1}(b_j) < 1$. We assume in addition that the sequence $(b_j)_{j\geq 1}$ belongs to $\ell^p(\mathcal{F})$ for some $p < 1$.

The optimization problem giving h_ν has a simple solution since the numbers ρ_j can be optimized separately. In view of the $\{\frac{r}{2}\}$-admissibility condition according to (1.4.11), it is easily checked that the optimization problem has a solution $\rho^* = (\rho_j^*)_{j\geq 1}$ where

$$\rho_j^* - 1 = \frac{r - \frac{r}{2}}{b_j}, \qquad j \geq 1, \tag{3.6.9}$$

and therefore

$$h_\nu := \frac{2\|f\|_{V^*}}{r} \prod_{j\geq 1} \Big(\frac{b_j}{b_j + \frac{r}{2}}\Big)^{\nu_j}. \tag{3.6.10}$$

Since the sequence $(h_\nu)_{\nu\in\mathcal{F}}$ is monotone decreasing, the best n-term sets Λ_n associated with the sequence $(h_\nu)_{\nu\in\mathcal{F}}$ can be viewed as the set of those indices ν such that h_ν exceeds a certain threshold $t = t(n) > 0$ that decreases with n, and are therefore of the form

$$\Lambda_n := \Big\{\nu \in \mathcal{F} \ : \ \sum_{j\geq 1} \alpha_j \nu_j \leq \Theta(n)\Big\} \ \text{ with } \ \alpha_j := \log\Big(\frac{b_j + \frac{r}{2}}{b_j}\Big) \ \text{ and } \ \Theta(n) = \log\Big(\frac{2\|f\|_{V^*}}{t(n) r}\Big)\,. \tag{3.6.11}$$

Note that in finite dimensional setting ($d < \infty$), if all b_j (and therefore α_j) were equal, then this would give the same a-priori simplex choice $\mathcal{S}_k$ defined in (3.6.3). In our case, the weight factors α_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. We should note that the weights α_j increase logarithmically in $\frac{1}{b_j}$

which is different than the initial intuitive choice (3.6.6) where it seemed, in view of the affine dependence $a(y) = 1 + \sum_j y_j b_j$, that the choice $\alpha_j = \frac{1}{b_j}$ is more judicious.

AJOUTER LA DESCRIPTION DE LA CONSTRUCTION DE L'ENSEMBLE

The general case

For a more general case, the computation of the values h_ν is not immediate. We show here that one can rely on computable sequence, yet preserving the rate of best n-term approximations.

First, we consider the setting where the number h_ν can be computed when needed at a unit cost. We then consider Algorithm 1.5.2 introduced in Chapter 1, namely

- Set $\Lambda_1 := \{0_{\mathcal{F}}\}$. For $n \geq 1$ do;
- Λ_n has been defined, compute $\mathcal{I}_1(\Lambda_n) := \{\nu \notin \Lambda_n : \nu - e_j \in \Lambda_n \text{ for any } j \text{ such } \nu_j \neq 0\}$;
- Get $\mu^n := \operatorname{argmax}_{\mu \in \mathcal{I}_1(\Lambda_n)}(h_\mu)$ and set $\Lambda_{n+1} := \Lambda_n \cup \{\mu^n\}$;

Although the immediate margin $\mathcal{I}_1(\Lambda)$ of any lower set Λ is of infinite cardinality (for $d = \infty$), the previous algorithm can be implemented in practice. One only needs to have a full knowledge of the numbers h_ν for ν in $\mathcal{I}_1(\{0_{\mathcal{F}}\}) = \{e_1, e_2, \ldots\}$ and has them sorted. We give in the following a justification of this observation.

We introduce first the notation $\operatorname{supp}(\Lambda)$ for support of Λ, which is the set (of integer) corresponding to all the directions activated by polynomial approximation in Λ, i.e.

$$\operatorname{supp}(\Lambda) := \bigcup_{\nu \in \Lambda} \operatorname{supp}(\nu). \tag{3.6.12}$$

We have the following lemma.

Lemma 3.6.1

Let Λ be a finite lower set, $\mu \in \mathcal{I}_1(\Lambda)$ and $\Lambda' = \Lambda \cup \{\mu\}$. Then the set $\mathcal{I}_1(\Lambda') \setminus \mathcal{I}_1(\Lambda)$ is finite and of cardinality at most $\#(\operatorname{supp}\Lambda) + 1$.

Proof: It is easily checked that $\mathcal{I}_1(\Lambda') \setminus \mathcal{I}_1(\Lambda) \subset \{\mu + e_j : j \geq 1\}$. This implies that

$$\mathcal{I}_1(\Lambda') \setminus \mathcal{I}_1(\Lambda) \subset \{\mu + e_j : j \in \operatorname{supp}(\Lambda')\},$$

Indeed, given $j \notin \operatorname{supp}(\Lambda')$ and assuming for example $\mu_1 \neq 0$, we have $\mu + e_j - e_1$ is not supported in $\operatorname{supp}(\Lambda')$ hence it is not in Λ' which implies that $\mu + e_j$ is not in $\mathcal{I}_1(\Lambda')$. The set $\mathcal{I}_1(\Lambda') \setminus \mathcal{I}_1(\Lambda)$ is then finite with cardinality smaller than $\#(\operatorname{supp}\Lambda') \leq \#(\operatorname{supp}\Lambda) + 1$. ■

Assuming the knowledge of all the values h_{e_j} for $j \geq 1$, the number of overall boundary value problems resolutions needed to obtain a final set Λ_{n^*} using the previous algorithm then does not exceed

$$\#(\Lambda_{n^*}) + \#(\mathrm{supp}(\Lambda_{n^*})) + 1 \leq 2\#(\Lambda_{n^*}), \tag{3.6.13}$$

hence the linearity in cost. In practice, assuming the knowledge of all the values $h_{e_1}, h_{e_2}, \ldots$, the algorithm can be executed as follows:

- Set $\Lambda_1 := \{0_{\mathcal{F}}\}$, then $\mathcal{S} := \mathcal{I}_1(\Lambda_1) = \{e_1, e_2, \ldots\}$ and sort it in a decreasing order comparing the values h_ν for $\nu \in \mathcal{S}$. For $n \geq 1$ do;
- Λ_n has been defined. Compute the finite set $\mathcal{I}_1(\Lambda_n) \setminus \mathcal{I}_1(\Lambda_{n-1})$ and merge it with $\mathcal{S}$ keeping the ordering;
- Get μ^n the first index in $\mathcal{S}$ and set $\Lambda_{n+1} = \Lambda_n \cup \{\mu^n\}$;

The algorithm might not be feasible since the numbers h_ν are sometime difficult to compute. The inspection of the proof of Theorems 1.4.5 and 1.6.5 show that there exist a computable sequence $(q_\nu)_{\nu\in\mathcal{F}}$ which is monotone decreasing and in $\ell^p(\mathcal{F})$ and bounds the sequence $(h_\nu)_{\nu\in\mathcal{F}}$. Indeed, we introduce the sequence $b := (b_j = \|\psi_j\|_{L^\infty})_{j\geq 1}$ and let J and κ defined by

$$\sum_{j>J} b_j \leq \frac{r}{8e}, \qquad \kappa = 1 + \frac{r}{8\|b\|_{\ell^1}} \tag{3.6.14}$$

We then define the sequence q_ν by $q_{0_{\mathcal{F}}} = 1$, then for $\nu \neq 0_{\mathcal{F}}$

$$q_\nu = q_E(\nu)q_F(\nu), \qquad q_E(\nu) := \prod_{j\leq J:\nu_j\neq 0} \kappa^{-\nu_j}, \qquad q_F(\nu) := \prod_{j>J:\nu_j\neq 0} \Big(e + \frac{r/4}{b_j}\frac{\nu_j}{|\nu_F|+1}\Big)^{-\nu_j}. \tag{3.6.15}$$

where $|\nu_F| = \sum_{j\geq J} |\nu_j|$ and $q_F(\nu)$ is equal to 1 when ν is supported in $\{1, \ldots, J\}$. By the arguments used in the proof of Theorem 1.6.5, the sequence $(q_\nu)_{\nu\in\mathcal{F}}$ is monotone decreasing and belongs to $\ell^p(\mathcal{F})$. Moreover, we have

$$h_\nu \leq \frac{2\|f\|_{V^*}}{r} q_\nu, \qquad \nu \in \mathcal{F}. \tag{3.6.16}$$

We can then use this computable sequence $(q_\nu)_{\nu\in\mathcal{F}}$ with the algorithm above. However we still need to address the problem of infinite cardinality of the first immediate margin $\mathcal{I}_1(\{0_{\mathcal{F}}\}) = \{e_1, e_2, \ldots\}$.

From the definition of the sequence $(q_\nu)_{\nu\in\mathcal{F}}$, we have

$$q_{e_j} = \kappa^{-1}, \quad j \leq J \quad \text{and} \quad q_{e_j} = \Big(e + \frac{r}{8b_j}\Big)\kappa^{-1}, j > J \tag{3.6.17}$$

From this, one can compute all the values q_{e_j}. However, one does not need to compute them all at once. Indeed, since for any $j > J$

$$q_{e_j} \leq \frac{8}{r} b_j = \frac{8}{r} \|\psi_j\|_{L^\infty} \to_{j\to\infty} 0, \tag{3.6.18}$$

then one can first activate the directions $e_1, \ldots, e_J$ and work in the finite dimensional setting with J, then activates (unlock) progressively the remaining directions $e_j, j > J$ whenever in the previous algorithm, the largest value in the sorted set $\mathcal{S}$ become smaller than the estimate $\frac{8}{r}\|\psi_j\|_{L^\infty}$.

We should note that all the previous considerations are irrelevant in finite dimension. However they can still be useful for reducing the numerical cost in the case where $d >> 1$.

3.6.2 Largest neighbor algorithm

Since the previous algorithm yields near optimal approximation, yet using a relatively simple greedy procedure, it is appealing to apply it directly with the Taylor energies. We consider the following algorithm

- Set $\Lambda_1 := \{0_{\mathcal{F}}\}$. For $n \geq 1$ do;
- Λ_n has been defined, compute $\mathcal{I}_1(\Lambda_n)$, the coefficients t_ν and their energies $c_\nu = \|t_\nu\|_{\overline{a}}$ for $\nu \in \mathcal{I}_1(\Lambda_n)$;
- Get $\nu^{n+1} = \text{argmax}_{\nu\in\mathcal{I}_1(\Lambda_n)}(c_\nu)$ and set $\Lambda_{n+1} = \Lambda_n \cup \{\nu^{n+1}\}$;

The intuition for considering such an algorithm is that if the sequence $(c_\nu)_{\nu\in\mathcal{F}}$ were monotone decreasing, then this would select the c_ν in decreasing order, yielding optimality in (3.2.12). In comparison with the bulk chase algorithms 3.3.1 and 3.4.3, the potential pay-off here is that the reduced margins $\mathcal{I}_1(\Lambda_n)$ are much smaller than $\mathcal{M}_n = \mathcal{M}(\Lambda_n)$. In addition, as we already proved in Lemma 3.6.1, at most $\#(\text{supp}(\Lambda_n)) + 1$ boundary value problems need to be solved at each step n. As we shall see in the numerical results section, this strategy gives excellent results. However, unlike Algorithms 3.3.1 and 3.4.3, we have no theoretical justification that it should perform optimally in the sense of convergence rates.

As explained in the previous section, the previous algorithm is feasible in infinite dimension whenever we have full knowledge of the coefficient $c_{e_j} = \|t_{e_j}\|_{\overline{a}}$ for every $j \geq 1$ or of their decays. Since the Taylor coefficients satisfy recursive formulas (3.2.3), we are able to retrieve a certain decay of the coefficients c_{e_j}. We have for any $j \geq 1$

$$\int_D \overline{a}(x)\nabla t_{e_j}(x)\nabla w(x) = -\int_D \psi_j(x)\nabla t_0(x)\nabla w(x), \qquad w \in V. \tag{3.6.19}$$

Setting $w = t_{e_j}$ and using Cauchy-Schwartz inequality, we obtain that

$$c_{e_j} \leq c_{0_{\mathcal{F}}} \| \frac{\psi_j}{\overline{a}} \|_{L^\infty(D)}. \tag{3.6.20}$$

Then as explained in the previous section, at step 1 we only compute the coefficients t_{e_j} and corresponding energies c_{e_j} for values $j = 1, \ldots, J$ with J some integer and then work as in dimension J. In the following steps n, we activate progressively the direction e_j for $j > J$ whenever the greatest value in the reduced margin $\mathcal{I}_1(\Lambda_n)$ is smaller than the estimates $c_{0_{\mathcal{F}}} \| \frac{\psi_j}{\overline{a}} \|_{L^\infty(D)}$.

3.6.3 Largest neighbor estimate algorithm

In order to save further computational cost, we can use majorants of c_ν in order to decide on the new set Λ_{n+1}. From (3.2.14), one straightforward upper estimate for c_ν is

$$c_\nu \leq E_\nu := \Big(\frac{1}{1+\gamma} \sum_{j:\nu_j \neq 0} \Big\| \frac{\psi_j}{\overline{a}} \Big\|_{L^\infty(D)} c^2_{\nu - e_j} \Big)^{\frac{1}{2}}, \qquad \gamma = \frac{r}{R}. \tag{3.6.21}$$

One can then construct the new set Λ_{n+1} as in the largest neighbour algorithm by using the estimates E_ν instead of c_ν. The saving comes from the fact that growing Λ_n into Λ_{n+1} does not require solving many boundary value problems. After each enrichment step n, one only needs to compute $t_{\nu^{n+1}}$ and $c_{\nu^{n+1}}$ for the new added index ν^{n+1} and then go to step $n+1$. In particular, one merely solve $J + \#(\Lambda_{n^*})$ boundary value problems in order to output a final set Λ_{n^*} where J is the number of directions $e_1, \ldots, e_J$ one choose to activate at step 1.

3.7 Numerical experiment

In this section, we study the numerical performance of the bulk chasing algorithms described in this chapter. For such algorithms, the choice of Λ_n is made adaptively and it is based on a bulk search procedure. We want to compare the effectiveness of this procedure when compared with non-adaptive strategies or the adaptive alternative strategies described in the previous section.

The algorithms that we analyzed in the previous sections were formulated in both the case where $d = \infty$ and $d < \infty$. In the present numerical tests, we use a parameter vector $y = (y_j)_{j=1,\cdots,d}$, of dimension d that ranges up to 255. More precisely, we consider on the unit square $D := [0,1] \times [0,1]$ the following problem,

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

where for illustration purposes we take $f(x) = f(x_1, x_2) := x_1 x_2$. As to the diffusion coefficient $a(x, y)$ and the choice of the ψ_j we consider two different test cases.

Test 1: characteristic functions. We partition D into 64 (8×8) squares D_j of equal shape and consider a diffusion coefficient that is piecewise constant on each subdomain:

$$a(x, y) = \overline{a}(x) + \sum_{j=1}^{64} y_j \psi_j(x), \quad \text{where } \overline{a} = 1 \text{ and } \psi_j = b_j \chi_{D_j}. \tag{3.7.1}$$

Since in this case the functions ψ_j have disjoint supports, the uniform ellipticity assumption $\mathbf{UEA}(r, R)$ simply means that the weights $b_j = \|\psi_j\|_{L^\infty(D)}$ are all strictly less than $1 - r$ for some number $r \in]0, 1[$. To study the consistency of the numerical results with our theory, we also require 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 ℓ^p-summable for some $0 < p < 1$. In our numerical test we take in (3.7.1)

$$b_j = \frac{0.9}{j^3}. \tag{3.7.2}$$

The uniform ellipticity assumption $\mathbf{UEA}(r, R)$ therefore holds with $r = 0.1$ and $R = 1.9$. This test is not physically realistic: it corresponds to a diffusion which is uncorrelated between the different subdomains and with variability that strongly differs between subdomains labelled by small and large values of j. Its main purpose is to compare adaptive strategies with a non-adaptive one based on the estimates (3.6.7), since these estimates can be explicitly computed for this test case as we explained in Section 3.6.1.

Test 2: wavelets. For the same domain $D = [0, 1] \times [0, 1]$, we consider the bidimensional Haar wavelet basis

$$h^i_{l,k}(x) = h^i(2^l x - k), \quad l \in \mathbb{N}, \quad k = (k_1, k_2) \in \{0, \ldots, 2^l - 1\}^2, \quad i = 1, 2, 3,$$

where the generating wavelets h^i are defined by

$$h^1(x_1, x_2) := \varphi(x_1) h(x_2), \quad h^2(x_1, x_2) := h(x_1) \varphi(x_2), \quad h^3(x_1, x_2) := h(x_1) h(x_2) \tag{3.7.3}$$

with

$$\varphi := \chi_{[0,1]} \quad \text{and} \quad h := \chi_{[0,1/2[} - \chi_{[1/2,1[} \tag{3.7.4}$$

Any function in $L^2(D)$ has a unique expansion into the orthogonal basis composed of all the above wavelets and of χ_D. We refer to [29] for a detailed treatment on wavelet bases. We consider a diffusion coefficient of the form

$$a(x, y) := \overline{a}(x) + \sum_{l=0}^{L} \beta_l \sum_{i=1}^{3} \sum_{k \in \{0, \ldots, 2^l - 1\}^2} y_{l,k,i} h^i_{l,k}(x), \tag{3.7.5}$$

where L is a fixed integer representing the finest scale level, $(\beta_l)_{l=0,\dots,L}$ a positive sequence and $y_{l,k,i}$ are the parametric variables ranging in $[-1,1]$. As in Test 1, we take $\overline{a}=1$.

With the above normalization $\|h^i_{l,k}\|_{L^\infty}=1$, it is known that the wavelet coefficients of a C^γ function decay like $\mathcal{O}(2^{-\gamma l})$ as the scale level grows. The rate of decay of the sequence β_l therefore reflects the amount of smoothness (or correlation in the stochastic context) in a. We consider the general form

$$\beta_l := c2^{-\gamma l} \quad c := 0.3\frac{2^\gamma-1}{1-2^{-L\gamma}}, \tag{3.7.6}$$

which ensures that the uniform ellipticity assumption $\mathbf{UEA}(r,R)$ holds with $r=0.1$ and $R=1.9$. In the numerical tests, we consider the two particular values

$$\gamma = 0.5 \ \text{ and } \ \gamma = 3,$$

in order to compare the effect of low smoothness (short range correlation) and of high smoothness (long range correlation) on the behaviour of our algorithms. Using the relabelling

$$\psi_j := \beta_l h^i_{l,k} \ \text{ and } \ y_j := y_{l,k,i}, \ \text{ when } \ j = 2^{2l}+3(2^lk_1+k_2)+i-1,$$

we may rewrite the above expansion (5.5.7) in the form $a(x,y) := \overline{a}(x)+\sum_{j=1}^d y_j\psi_j(x)$ adopted in this paper, with

$$d := 2^{2(L+1)}-1.$$

In order to study the robustness of the method to the dimensionality, we consider different values $L=1,2,3$ for the maximal scale, which corresponds to taking $d=15,63,255$. Note that after this relabelling, the sequence $(\|\psi_j\|_{L^\infty})_{j\geq 0}$ decays like $\mathcal{O}(j^{-\gamma/2})$, and therefore like $\mathcal{O}(j^{-1/4})$ and $\mathcal{O}(j^{-3/2})$ for $\gamma=0.5$ and $\gamma=3$ respectively.

As we considered in §3.5, we use one fixed finite element space for the spatial discretization of all the 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*

$$\Big\|u_h-\sum_{\nu\in\Lambda_n}t_{\nu,h}y^\nu\Big\|_{\mathcal{V}_\infty} := \sup_{y\in U}\Big\|u_h(y)-\sum_{\nu\in\Lambda_n}t_{\nu,h}y^\nu\ \Big\|_V, \tag{3.7.7}$$

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 (3.5.10).

We compare the three non-adaptive strategies (i) **QN** when $\Lambda_k=\mathcal{B}_k$ (ii) **PN** when $\Lambda_k=\mathcal{S}_k$ where $\mathcal{B}_k$ and $\mathcal{S}_k$ are defined in (3.6.3), (ii) **LE** when are Λ_k generated by the largest estimate algorithm described in Section 3.6.1, and the three adaptive strategies (iv) **BS** when Λ_k generated by bulk chase algorithm (v) **LN** when Λ_k generated by largest neighbour algorithm (vi) **LNE** when Λ_k generated by largest estimate neighbour algorithm

3.7.1 Numerical results for Test 1

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 3.7.1 the error curves for the six strategies described above for the generation of the sets Λ_n. These error curves represent the supremum error (3.7.7) (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 shows that the sets Λ_n selected by the adaptive algorithms change very little as we modify the spatial discretization. This suggests that the same sets and error curves would be obtained if there were no spatial discretization at all, i.e. if we were computing the t_ν by exactly solving the boundary value problems (3.2.3). 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.

- 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

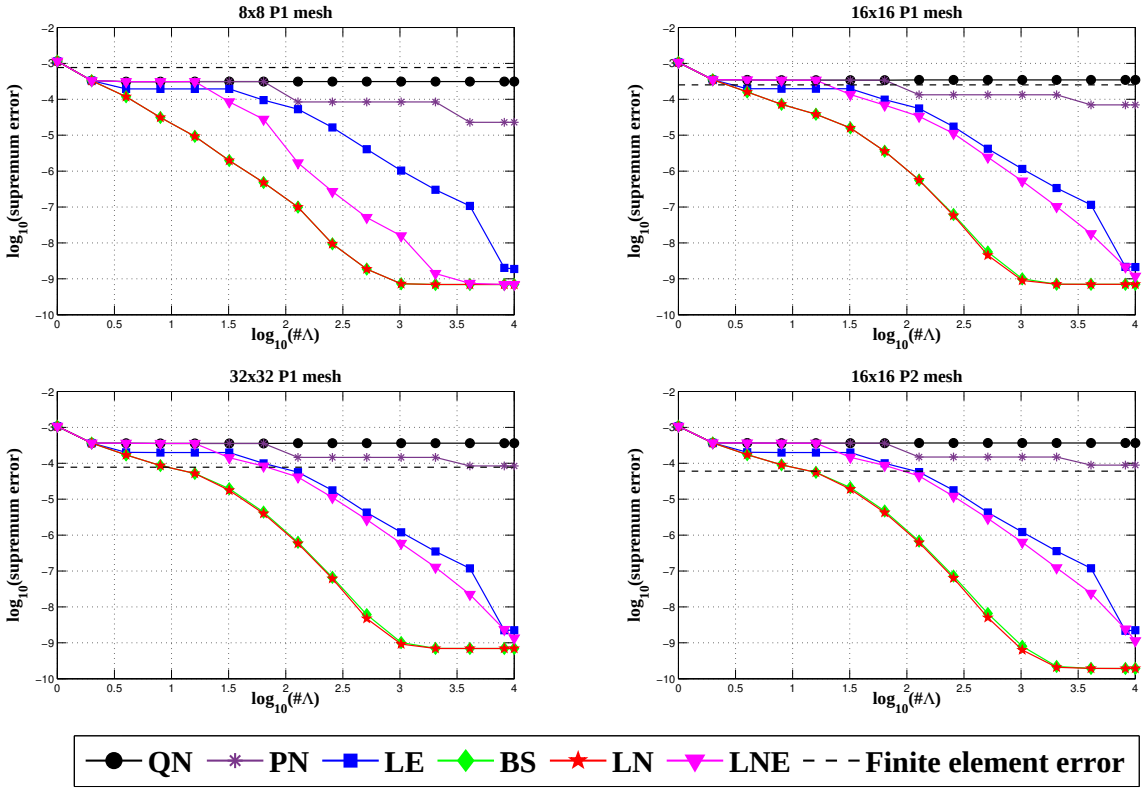

Figure 3.7.1: Comparison the different strategies for finite element spaces (i) (upper left), (ii) (upper right), (iii) (lower left) and (iv) (lower right).

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 3.7.2 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.

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

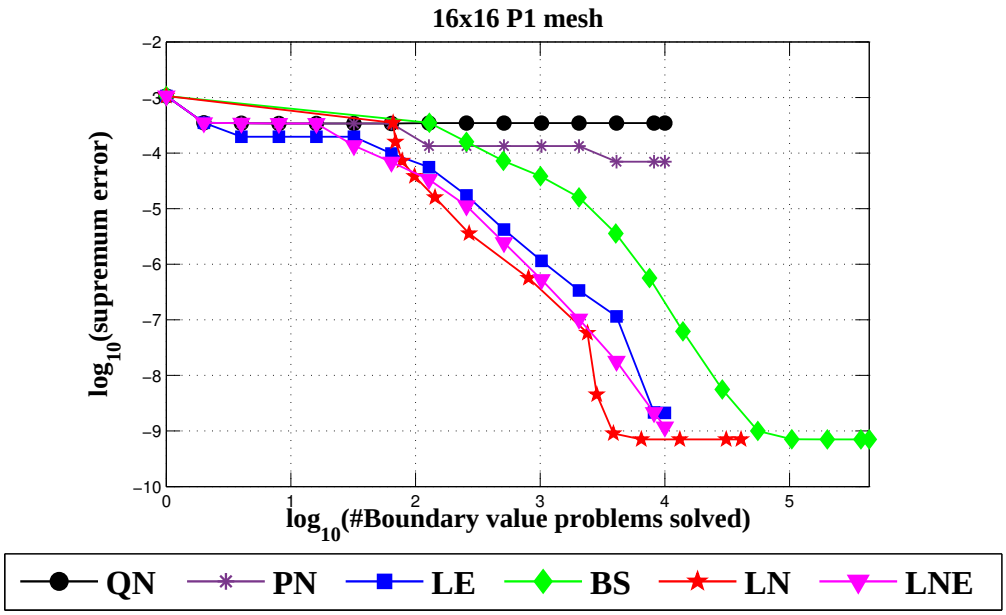

Figure 3.7.2: Comparison of the different strategies in term of total number of solved bvp

curves to estimate the number of Taylor coefficients that we need to compute with the fine discretization in order to reach a prescribed accuracy.

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

$$\sup_{y\in U} \|u_h(y) - \sum_{\nu\in\Lambda_n} t_{\nu,h} y^\nu\|_V \leq \varepsilon, \tag{3.7.8}$$

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

$$\sup_{y\in U} \|u_h(y) - u(y)\|_V. \tag{3.7.9}$$

While this last quantity is not exactly known to us, it can be bounded according to a-priori estimate (3.5.10) based on our knowledge of the maximal Sobolev smoothness of $u(y)$, or estimated in a finer way based on a-posteriori analysis.

In all three 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 strongly suggests that the presented algorithms will perform well also for more general parameter dependences, where the most significant coordinate appears only in high dimension.

We also have 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 3.7.3. For the MC method, we display the average of the error curves for 6 independent realizations in order to illustrate the 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, however with a different stiffness matrix for each problem.

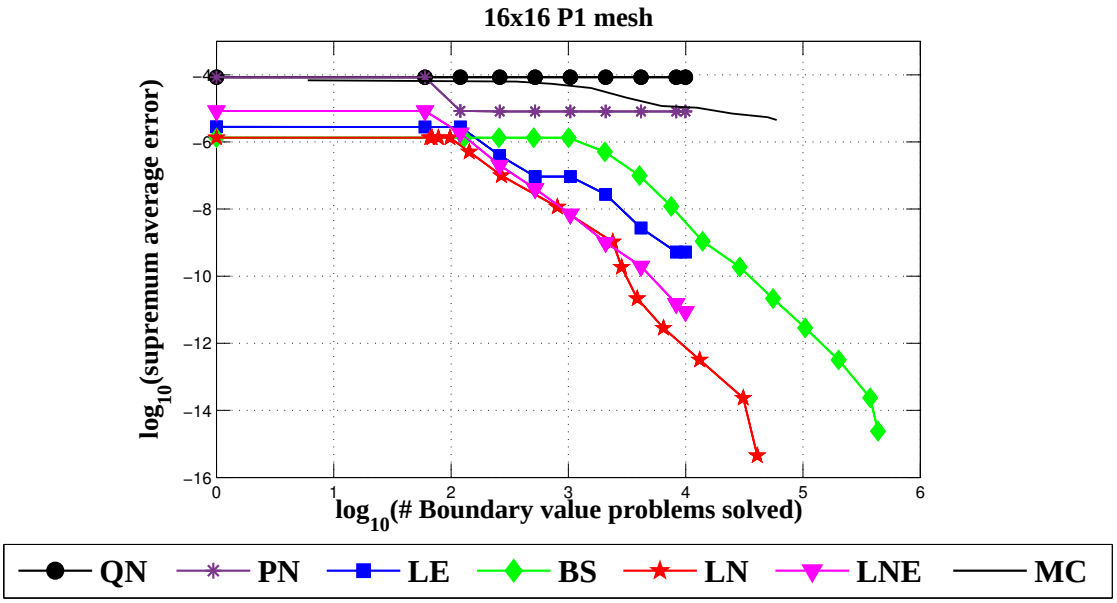

Figure 3.7.3: Comparison of the different strategies with Monte Carlo method.

3.7.2 Numerical results for Test 2

For this test, we only display the comparison between Algorithms PN, QN, LE, BS and LN, since we observed that the LNE algorithm does not perform as good as LN, similar to Test 1. In contrast to Test 1, we cannot use the a-priori bound

$$h_\nu := \frac{\|f\|_{V^*}}{\delta} \inf\{\rho^{-\nu} : \rho \text{ is } \delta\text{-admissible}\}. \tag{3.7.10}$$

with $\delta = r/2 = 0.05$ to derive the choice of the active indices for the largest estimates algorithm, since this optimization problem has no simply computable solution. We therefore rely on sub-optimal bounds of the form

$$h_\nu \leq q_\nu := \frac{2\|f\|_{V^*}}{r}\rho^{-\nu}, \tag{3.7.11}$$

where $\rho = \rho(\nu)$ is a particular sequence that is $\{\frac{r}{2}\}$-admissible, which is chosen depending on ν, in contrast to Test 1. Our best results were obtained with the following choice for $\rho(\nu)$:

$$\rho_j(\nu) = \frac{A(\nu)}{\|\psi_j\|_{L^\infty(D)}} \quad \text{if} \quad \nu_j \neq 0 \quad \text{and} \quad \rho_j(\nu) = 0 \quad \text{if} \quad \nu_j = 0, \tag{3.7.12}$$

where $A(\nu)$ is the largest positive number such that the result $\rho(\nu)$ is $\{\frac{r}{2}\}$-admissible. This number is thus defined in such a way that

$$\left\| \sum_{j \text{ s.t. } \nu_j \neq 0} \rho_j(\nu) |\psi_j| \right\|_{L^\infty} = 1 - \frac{r}{2} = 0.95.$$

The sets Λ_n chosen in Algorithm LE now correspond to the n largest q_ν.

Figure 3.7.4 displays the error curves in terms of the number of solved boundary value problems, similar to Figure 3.7.2 for Test 1, for the two values $\gamma = 0.5$ and $\gamma = 3$, and for different maximal scale level $L = 1, 2, 3$, corresponding to dimensions $d = 15, 63, 255$. We record several observations about the error curves:

- In the low smoothness/correlation case $\gamma = 0.5$, Algorithms BS, LE and LN do not perform significantly better than Algorithm PN which corresponds to the standard choice of polynomials of fixed total degree. This can be explained by the fact that, in this case, all $\|\psi_j\|_{L^\infty}$ have roughly the same range of magnitude so that all variables y_j are equally active. In turn, the active index sets selected by BS, LE and LN are not highly anisotropic, and perform similar than those of PN. In contrast, the high smoothness/correlation case $\gamma = 3$ highly benefits from an anisotropic selection (higher degree tends to be allocated to the variables associated to coarse scale wavelets), and in turn Algorithms BS, LE and LN significantly outperform PN and QN.

- In the low smoothness/correlation case $\gamma = 0.5$, all algorithms are subject to the curse of dimensionality in the sense that the error curves deteriorate as d increases. This includes the adaptive algorithms, which is not in contradiction with our theoretical results. Indeed, this value of γ corresponds to a decay in $\mathcal{O}(j^{-1/4})$ for the sequence $(\|\psi_j\|_{L^\infty})_{j\geq 0}$, which is therefore not ℓ^p summable for any value $p < 1$. In contrast, the high smoothness/correlation case $\gamma = 3$ is not subject to the curse of dimensionality when Algorithms LE and LN are being used. Note that this value of γ corresponds to a decay in $\mathcal{O}(j^{-3/2})$ for the sequence $(\|\psi_j\|_{L^\infty})_{j\geq 0}$, which is therefore ℓ^p summable for $p > 2/3$.

- For both values of γ, Algorithms LN and LE gives the best performances, and Algorithm BS is subject to the curse of dimensionality due to the cost of solving boundary value problems for all indices in the margin $\mathcal{M}_n$.

- The error curves for Algorithm LN and BS start decreasing only after a certain number of boundary value problems has been solved. This is simply due to the fact that at the very first step, d boundary value problems have to be solved, corresponding to the cardinality of the margin $\mathcal{M}_0$ of $\Lambda_0 = \{0\}$ (which at this stage is the same as the reduced margin).

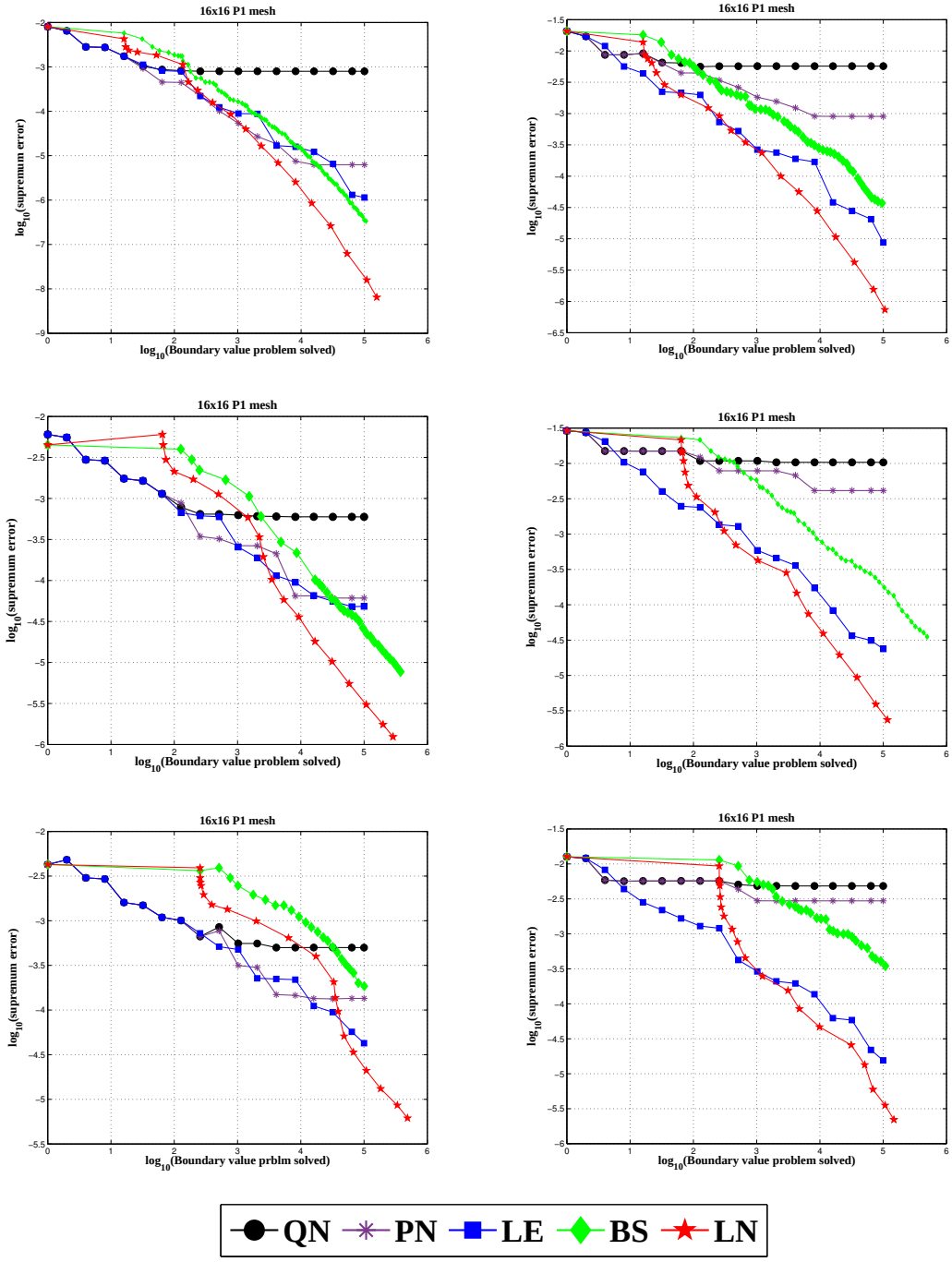

Figure 3.7.4: Comparison the different strategies for for $\gamma = 0.5$ (left) and $\gamma = 3$ (right), and different dimensionalities $d = 15$ (up), $d = 63$ (middle) and $d = 255$ (bottom).

In order to have an idea of the geometry of the coefficients sets Λ produced by the different strategies, consider the projection on two variables $j = 1, 5$, associated to the

first and second scale level $l = 0, 1$, i.e. the sets

$$\{(\nu_1, \nu_5) \ : \ \nu \in \Lambda\}.$$

We compare these sets on Figure 3.7.5, when $\#(\Lambda) = 10000$ for the strategies QN, PN, LE and LN in the case $\gamma = 3$ and $d = 63$. 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 $d = 63$ the dimension of the spaces $\mathbb{P}_{\mathcal{B}_1}$ of polynomials of degree at most 1 in each variable and $\mathbb{P}_{\mathcal{S}_3}$ of total degree 3 clearly exceeds 10000 and therefore no degree higher than 1 and 3 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 higher polynomial degrees in the most active variable y_1. We did not plot the sets generated by BS which are quite similar to LN. Note however that their geometry differs from that of the set generated by LE based on the a-priori estimates q_ν.

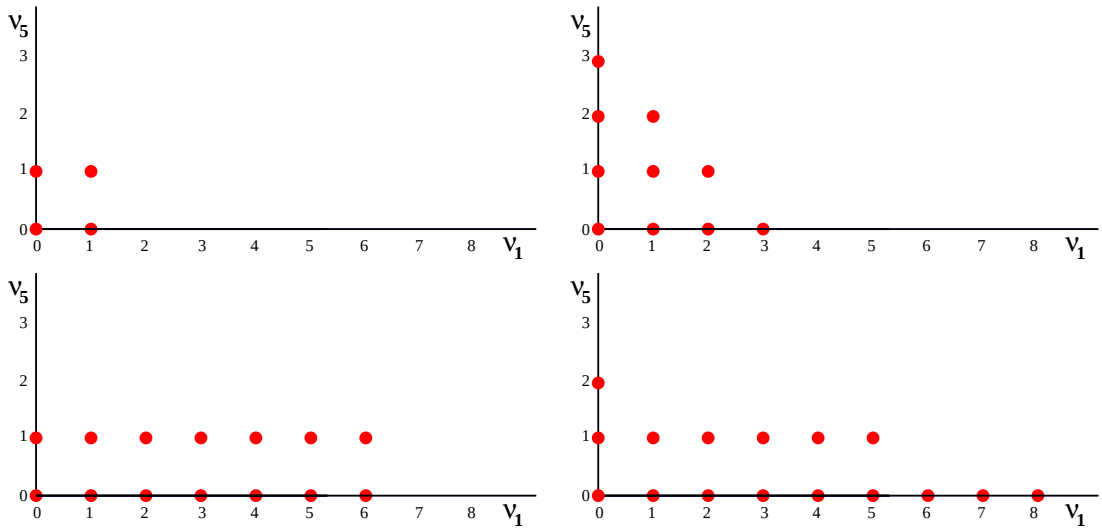

Figure 3.7.5: Comparison the index sets Λ with $\#(\Lambda) = 10000$ projected on the componetry $(1, 5)$ for Algorithms QN (upper left), PN (upper right), LE (lower left) and LN (lower right).

3.8 Conclusion

In this chapter, we have introduced adaptive algorithms for sparse Taylor approximation of parametric and stochastic PDEs. These algorithms have several remarkable features, in particular:

(i) They build the polynomial expansion by solving a sequence of boundary value problems (3.2.3) which all have the same stiffness matrix.

(ii) From a theoretical point of view, their convergence with respect to the polynomial dimension can be proved to be near optimal.

This second property is reflected in the numerical tests. In contrast to other approaches for selecting the active index sets, the algorithm does not use any information based on a-priori analysis, and yet performs at least as good as when using such a-priori choices.

It is worth mentioning that this approach can be applied verbatim to other models, such as a parabolic equation of the form

$$\partial_t u - \operatorname{div}(a\nabla u) = f, \quad \text{in } [0,T]\times D, \quad u(x,0) = u_0(x), \quad u(x,t) = 0, \ x \in \partial D, \quad (3.8.1)$$

with a of the same parametric form as in this chapter and the functions $\psi_j = \psi_j(x,t)$ satisfying a similar condition as $\mathbf{UEA}(r,R)$. In that case, the solution space is $V = L^2([0,T],H_0^1(D))$. However, let us stress that our approach is strongly tied to the affine structure of a with respect to the parameter vector y, in contrast to other methods such as collocation.

In the next chapter, we investigate the approximation in the mean square sense. We have proved in chapter 1 that approximation of u by Legendre series truncated to their best n-terms converge in this sense with the rate $(n+1)^{-s^*}$ with $s^* = \frac{1}{p} - \frac{1}{2}$. Our goal is then to retrieve this rate through practical algorithms.

Chapter 4

An adaptive algorithm for sparse Galerkin approximations

Contents

4.1 Introduction

In this chapter, we study the second intrusive method for the approximation of the parametric elliptic model of Chapter 1. The model is given by the equation (1.1.1)

where the diffusion coefficient a depends on the parameter y in an affine manner as in (1.1.2) and satisfies the uniform ellipticity assumption $\mathbf{UEA}(r, R)$ given in (1.1.3). In the present setting, y is a random vector with joint probability ϱ and one is then interested in approximating the solution map

$$y \in [-1,1]^{\mathbb{N}} \mapsto u(y) \in H_0^1(D), \tag{4.1.1}$$

in the mean square sense, i.e. in the Bochner space

$$\mathcal{V}_2 := L^2(U, V, d\varrho), \quad \text{where} \quad U := [-1,1]^{\mathbb{N}} \quad \text{and} \quad V := H_0^1(D). \tag{4.1.2}$$

The diffusion coefficient a is a random field on a probability space (Ω, Σ, P) over $L^\infty(D)$ (see, e.g., [61]) and the right hand side f is a given non random function on D. In this model, the random parameter $y := (y_j)_{j\geq 1}$ is used to describe the uncertainty in the diffusion coefficient a, through Karhunen-Loève expansion for example, see the discussion in the general introduction.

Rather than striving for at most generality, we consider that the random variables y_j are independent and identically distributed with respect to the uniform measure in $[-1,1]$. Therefore ϱ, the joint probability distribution of the random vector y, is the uniform probably measure over U

$$d\varrho(y) := \otimes_{j\geq 1} \frac{dy_j}{2}. \tag{4.1.3}$$

Let us observe that

$$\|v\|_{\mathcal{V}_2} := \Big(\int_U \|v(y)\|_V^2 d\varrho(y) \Big)^{\frac{1}{2}} = \Big(\int_U \int_D |\nabla v(y)|^2 dx d\varrho(y) \Big)^{\frac{1}{2}}, \tag{4.1.4}$$

induces on $\mathcal{V}_2$ the structure of a separable Hilbert space equipped with the inner product

$$\langle v, w \rangle := \int_U \int_D \nabla v(y) \nabla w(y) dx d\varrho(y). \tag{4.1.5}$$

As in chapter 3, we recall the theoretical approximation results of the solution map u, which is here a random field over V, in the space $\mathcal{V}_2$. In chapter 1, it is shown that a similar theorem of 3.1.1 holds with respect to tensor product Legendre expansions. We state the result which is a benchmarks for Galerkin approximations discussed in this chapter.

We use the same notations of the previous chapters. We consider $\mathcal{F}$ the (countable) set of all sequences of nonnegative integers which are finitely supported. We consider the Legendre expansions of the solution map u introduced in §1.2 of Chapter 1. Since $u \in \mathcal{V}_\infty \subset \mathcal{V}_2$, it admits unique expansions

$$u(y) = \sum_{\nu \in \mathcal{F}} v_\nu L_\nu = \sum_{\nu \in \mathcal{F}} u_\nu P_\nu, \tag{4.1.6}$$

where

$$v_\nu := \int_U u(y) L_\nu(y) d\varrho(y) \in V \quad \text{and} \quad u_\nu := v_\nu \prod_{j\geq 1} \sqrt{1+2\nu_j}. \tag{4.1.7}$$

The following theorem is the analog to Theorem 3.1.1 for Taylor expansions.

Theorem 4.1.1

If the sequence $b := (\|\psi_j\|_{L^\infty})_{j\geq 1}$ belongs to $\ell^p(\mathbb{N})$ for some $0 < p < 1$, then the sequences $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$.

Theorem 4.1.1 has certain implications on the approximation of the solution u in the infinite and mean square senses by Legendre series. We denote by $(\Lambda_n^L)_{n\geq 1}$ and $(\Lambda_n^P)_{n\geq 1}$ sequences of nested sets of indices $\nu \in \mathcal{F}$ corresponding to the n largest values of $\|v_\nu\|_V$ or $\|u_\nu\|_V$ respectively, We have the convergence estimate

$$\Big\|u - \sum_{\nu\in\Lambda_n^L} v_\nu L_\nu\Big\|_{\mathcal{V}_2} = \Big(\sum_{\nu\notin\Lambda_n^L} \|v_\nu\|_V^2\Big)^{\frac{1}{2}} \leq \Big\|(\|v_\nu\|_{\mathcal{V}_2})\Big\|_{\ell^p(\mathcal{F})} (n+1)^{-s^*}, \quad s^* := \frac{1}{p} - \frac{1}{2}. \tag{4.1.8}$$

and

$$\Big\|u - \sum_{\nu\in\Lambda_n^P} u_\nu P_\nu\Big\|_{\mathcal{V}_\infty} \leq \Big\|(\|u_\nu\|_V)\Big\|_{\ell^p(\mathcal{F})} (n+1)^{-s}, \quad s := \frac{1}{p} - 1. \tag{4.1.9}$$

Considering rather the monotone envelopes of the sequences $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$, see the definition in (1.5.4), which are monotone decreasing, we can localize the best n-term index sets to lower sets (by considering lower realizations, see Definiton 1.5.3), yet preserving the same decay rate. Namely, if $(\Lambda_n^{L*})_{n\geq 1}$ and $(\Lambda_n^{P*})_{n\geq 1}$ are sequences of nested lower sets associated with $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$ and $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$, then

$$\Big\|u - \sum_{\nu\in\Lambda_n^{L*}} v_\nu L_\nu\Big\|_{\mathcal{V}_2} = \Big(\sum_{\nu\notin\Lambda_n^{L*}} \|v_\nu\|_V^2\Big)^{\frac{1}{2}} \leq \Big\|(\|v_\nu\|_{\mathcal{V}_2})\Big\|_{\ell^p_m(\mathcal{F})} (n+1)^{-s^*}, \tag{4.1.10}$$

and

$$\Big\|u - \sum_{\nu\in\Lambda_n^{P*}} u_\nu P_\nu\Big\|_{\mathcal{V}_\infty} \leq \Big\|(\|u_\nu\|_V)\Big\|_{\ell^p_m(\mathcal{F})} (n+1)^{-s}, \tag{4.1.11}$$

with s and s^* as above. We are able then to approximate simultaneously all the functions of the solution manifold

$$\mathcal{M} := \Big\{u(y)\ ;\ y \in U\Big\}, \tag{4.1.12}$$

at the cost of computing n coefficients $u_\nu \in V$ with the rate $(n+1)^{-s}$ which is as good the rate one can get by truncated Taylor series. Moreover, Legendre series may also provide approximation with the improved rate $(n+1)^{-s^*}$, $s^* = s + \frac{1}{2}$, in the mean square sense.

However, even if the index sets $\Lambda_n^{L^*}$ for example are known, the associated Legendre series are computationally out of reach. Indeed, in contrast to Taylor coefficients, the

affine dependance of the diffusion coefficient a on y does not yield a simple recursion for the computation of Legendre coefficients. It is not even explicit how to compute the Legendre coefficient

$$v_{0_{\mathcal{F}}} := \int_U u(y) d\varrho(y) = \mathbb{E}[u(y)], \tag{4.1.13}$$

associated with the polynomial $L_{0_{\mathcal{F}}} \equiv 1$.The rates in (4.1.8), (4.1.10), (4.1.9) and (4.1.11) should then only be considered as benchmark rates. Near optimal computable approximations are to be investigated. We focus our efforts in this direction in the sense of (4.1.8) and (4.1.10) since we already have satisfactory results for approximations in the sense of (4.1.9) and (4.1.11) by computable Taylor series. In others words, we only target the approximation of the solution map u in the mean square sense with convergence rate $(n+1)^{-s^*}$.

The goal in this chapter is to give concrete algorithms that adaptively build near optimal sequences $(\Lambda_k)_{k\geq 0}$, at costs that scales linearly in $\#(\Lambda_k)$, and corresponding Galerkin projection $u_{\Lambda_k} \in \mathbb{V}_{\Lambda_k} = V \otimes \mathbb{P}_{\Lambda_k}$ which converge toward u with a rate $(n+1)^{-s^*}$ where $n = n(k) = \#(\Lambda_k)$. The techniques developed for this purpose are, as in Chapter 3, the adaptive strategies for wavelet methods [30, 31, 49] or for finite element methods [46, 67, 13, 78]. In particular, we use a *bulk chasing procedure* in order to build the set Λ_{k+1} knowing the set Λ_k.

The outline of this chapter is similar to that of Chapter 3. We first simplify the problem into a residual reduction problem meeting to some extent the framework investigated in [30, 31, 49], then we show, as in Chapter 3, that the affine parametric dependence (1.1.2) has numerically useful implications on the reduction of the residuals, and finally propose algorithms that exploit such implications and yield near optimal convergence. As in Chapter 3, we discuss the feasibility of the algorithms in the infinite dimension setting $d = \infty$ and the effects of numerical discretization.

In §4.2, using Legendre polynomials we show that the problem can be reformulated in the more convenient problem of finding near optimal Galerkin projection associated to an infinite system

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{4.1.14}$$

where $\mathbf{A}$ is an infinite matrix of operators from V into V^* and $\mathbf{u}$ and $\mathbf{f}$ are merely the vectors of Legendre coefficients of the solution map u and the source term f in the Legendre basis. We then establish some properties of the matrix $\mathbf{A}$ which are necessary to the analysis of the bulk chasing algorithm discussed later. We should note however, that unlike in [30, 31, 49] where $\mathbf{A}$ is an infinite real matrix which exhibits a fast decay away from its diagonal, here $\mathbf{A}$ is a matrix of operators which we can explicitly compute. However the key points of the analysis are the same.

In §4.3, We show that the residual $\mathbf{r}_\Lambda$ associated with a Galerkin projection $\mathbf{u}_\Lambda$ associated with (4.1.14) and an index set $\Lambda \subset \mathcal{F}$ is supported in a neighbourhood of Λ which, in the case where Λ is lower, coincides with the margin $\mathcal{M}(\Lambda)$ of Λ, defined in

Chapter 3. Using this property, we show how the residual can be reduced by growing Λ in the mentioned neighbourhood. Then, using the $\ell^p(\mathcal{F})$ summability of Legendre coefficients, we show that the cardinality of the enriching set can be controlled, leading as in Chapter 3 to provable convergence results. Finally, using the stronger $\ell^p_m(\mathcal{F})$ summability of Legendre coefficients, we show that all the previous results can be localized to lower sets.

Based on the reduction and cardinality controlability results found in §4.3, we propose in §4.4 two adaptive algorithms and prove that the index sets Λ_k generated by the algorithms are near optimal in the sense of (4.1.8) and (4.1.10). In other words, the associated Galekin projections converge toward u in the mean square sense with the optimal rate $(n+1)^{-s^*}$ with $n = n(k) = \#(\Lambda_k)$. As in Chapter 3, the algorithms studied can not be implemented in the infinite dimension setting and are costly in the setting $d >> 1$.

In §4.5, we remedy this defect by introducing a second algorithm which operates at step k the bulk search on restricted neighbourhoods of Λ_k obtained incrementally and which are of moderate cardinality even in the case $d = \infty$. We prove that this new realistic algorithm generates also index sets that are near optimal in the sense of (4.1.8) and (4.1.10).

In §4.6, we study 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 on D, for example by a finite element method on D. We prove that the additional error introduced by the finite element discretization is independent of the number of computed value problems.

Unlike Chapter 3 where Taylor series are computed exactly (or with controlled discretization error) once an index set Λ is considered, the Galerkin projections can only be approximated. We show in §4.7 that this can be done to any accuracy by an iterative Jacobi method and then propose a bulk chasing algorithm that takes into account this limitation, yet converges with a similar rate as the idealized algorithms.

In §4.8, we show that Galerkin projection can also be used to approximate the solution map u in the uniform sense. Using a growth result on certain quadratic sum of Legendre polynomials infinite norms, we show that the rate of convergence of Galerkin projections using lower sets is at worse deteriorated by $(n+1)$ when the uniform sense is considered.

4.2 Galerkin Approximations

4.2.1 Weak formulation

The solution map u belongs to $\mathcal{V}_2$ and can be defined as the unique solution of the variational problem

$$\mathcal{B}(u,v) = \mathcal{L}(v), \qquad v \in \mathcal{V}_2, \tag{4.2.1}$$

where we have defined the bilinear form $\mathcal{B}$ over $\mathcal{V}_2 \times \mathcal{V}_2$ and the linear form $\mathcal{L}$ over $\mathcal{V}_2$ by

$$\mathcal{B}(w,v) := \int_U \int_D a(x,y)\nabla w(x,y)\cdot\nabla v(x,y)dxd\varrho(y), \qquad \mathcal{L}(v) := \int_U \int_D f(x)v(x,y)dxd\varrho(y). \tag{4.2.2}$$

The integrals over D are understood as extensions by continuity from $L^2(D) \times L^2(D)$ to duality pairings between V^* and V. The bilinear form $\mathcal{B}$ is clearly symmetric. Moreover, the uniform ellipticity assumption $\mathbf{UEA}(r,R)$ implies that $\mathcal{B}$ is coercive and continuous. We denote by $\|\cdot\|_E$ the norm induced by $\mathcal{B}$ on $\mathcal{V}_2$, i.e.

$$\|v\|_E := \sqrt{\mathcal{B}(v,v)}, \qquad v \in \mathcal{V}_2. \tag{4.2.3}$$

This norm is then equivalent to the norm $\|\cdot\|_{\mathcal{V}_2}$ with

$$\sqrt{r}\|v\|_{\mathcal{V}_2} \le \|v\|_E \le \sqrt{R}\|v\|_{\mathcal{V}_2}, \qquad v \in \mathcal{V}_2. \tag{4.2.4}$$

For any index set $\Lambda \subset \mathcal{F}$, we define the polynomials space

$$\mathbb{V}_\Lambda(L) := \left\{ \sum_{\nu\in\Lambda} v_\nu L_\nu : v_\nu \in V \right\} = V \otimes \operatorname{span}\Big\{ L_\nu, \nu \in \Lambda \Big\}, \tag{4.2.5}$$

It readily seen that this space coincides with the polynomials space $\mathbb{V}_\Lambda := V \otimes \mathbb{P}_\Lambda$ if Λ is lower. Since we shall only work with Legendre polynomials and for the sake of notational clearness, even when Λ is not lower, we drop L from the notation $\mathbb{V}_\Lambda(L)$. We denote

$$u_\Lambda := \sum_{\nu\in\Lambda} u_{\Lambda,\nu} L_\nu \in \mathbb{V}_\Lambda \tag{4.2.6}$$

the Galerkin approximation of u in the space $\mathbb{V}_\Lambda$ with respect to the weak formulation (4.2.1), i.e. the unique solution to the variational problem

$$\mathcal{B}(u_\Lambda, v_\Lambda) = \mathcal{L}(v_\Lambda), \qquad v_\Lambda \in \mathbb{V}_\Lambda. \tag{4.2.7}$$

The computation of the Galerkin approximation requires a spacial discretization on the space variable x. We postpone this discussion to §4.6 and focus our analysis on the

approximation properties in y. First, the Galerkin approximation u_Λ is the projection of u with respect to the norm $\|\cdot\|_E$, hence it is optimal in the sense

$$\|u - u_\Lambda\|_E \leq \inf_{v_\Lambda \in \mathbb{V}_\Lambda} \|u - v_\Lambda\|_E. \tag{4.2.8}$$

The norm equivalency inequality (4.2.4) implies

$$\|u - u_\Lambda\|_{\mathcal{V}_2} \leq \sqrt{\frac{R}{r}} \inf_{v_\Lambda \in \mathbb{V}_\Lambda} \|u - v_\Lambda\|_{\mathcal{V}_2} = \sqrt{\frac{R}{r}} \Big\| u - \sum_{\nu \in \Lambda} u_\nu L_\nu \Big\|_{\mathcal{V}_2}. \tag{4.2.9}$$

This shows that Galerkin approximations can provide approximations to u with similar rates to Legendre series, up to a multiplicative constant factor $\sqrt{\frac{R}{r}}$. In particular, the Galerkin approximations $u_{\Lambda_n^L}$ and $u_{\Lambda_n^{L*}}$ where the sequences $(\Lambda_n^L)_{k\geq 1}$ and $(\Lambda_n^{L*})_{k\geq 1}$ are used in (4.1.8) and (4.1.10) are near optimal in the sense of the best n-term approximation. Unfortunately the construction of the sequences $(\Lambda_k^L)_{k\geq 1}$ and $(\Lambda_k^{L*})_{k\geq 1}$ suppose the full knowledge of the sequence $(\|v_\nu\|_V)_{\nu\in\mathcal{F}}$, which is the primary obstruction we intend to avoid. To overcome this drawback, we will rely on adaptive algorithms in the construction of the Galerkin approximations.

4.2.2 Sequence Space Reformulation

We introduce $\mathcal{A}$ the linear operator form $\mathcal{V}_2 = \mathcal{L}(U, V, d\varrho)$ into its dual $\mathcal{V}_2^*$ that associates to each v, $\mathcal{A}v$ defined by

$$\langle \mathcal{A}v, w\rangle = \mathcal{B}(v, w), \quad w \in \mathcal{V}_2, \tag{4.2.10}$$

where the scalar product $\langle\cdot,\cdot\rangle$ is the duality pairing between $\mathcal{V}_2^*$ and $\mathcal{V}_2$. The operator $\mathcal{A}$ is bounded and invertible. Now, since $f \in V^* \subset \mathcal{V}_2^*$, then the variational problem (4.2.1) satisfied by u is equivalent to $\langle \mathcal{A}u, v\rangle = \langle f, v\rangle$ for every $v \in \mathcal{V}_2$ or equivalently

$$\mathcal{A}u = f \quad \text{in} \quad \mathcal{V}_2^*. \tag{4.2.11}$$

We are interested in representing the previous system in the Legendre basis $(L_\nu)_{\nu\in\mathcal{F}}$. To this end, we introduce the Hilbert space

$$\ell^2(\mathcal{F}, V) := \Big\{ \mathbf{w} := (\mathbf{w}_\nu)_{\nu\in\mathcal{F}} \in V^{\mathcal{F}} : \|\mathbf{w}\|^2_{\ell^2(\mathcal{F},V)} := \sum_{\nu\in\mathcal{F}} \|\mathbf{w}_\nu\|_V^2 < \infty \Big\}, \tag{4.2.12}$$

with the inner product induced by $\|\cdot\|_{\ell^2(\mathcal{F},V)}$ defined in the obvious way. Since the family $(L_\nu)_{\nu\in\mathcal{F}}$ is an orthonormal basis of $\mathcal{V}_2$, we may define an isometry between the Hilbert spaces $\mathcal{V}_2$ and $\ell^2(\mathcal{F}, V)$ that associates to each $w \in \mathcal{V}_2$, the vector indexed by $\mathcal{F}$ of its Legendre coefficients $\mathbf{w} := (\mathbf{w}_\nu)_{\nu\in\mathcal{F}} \in \ell^2(\mathcal{F}, V)$. Now, we introduce the space

$$\ell^2(\mathcal{F}, V^*) := \{ \mathbf{W} := (\mathbf{W}_\nu)_{\nu\in\mathcal{F}} \in V^{*\mathcal{F}} : \|\mathbf{W}\|^2_{\ell^2(\mathcal{F},V^*)} := \sum_{\nu\in\mathcal{F}} \|\mathbf{W}_\nu\|_{V^*}^2 < \infty \}, \tag{4.2.13}$$

Since $\mathcal{V}_2^* = (L^2(U,V,d\varrho))^* \simeq L^2(U,V^*,d\varrho)$, we may also define in the same way an isometry between the Hilbert space $\mathcal{V}_2^*$ and $\ell^2(\mathcal{F},V^*)$. To be consistent with the notation of the elements of the space $\ell^2(\mathcal{F},V^*)$, we introduce the sequence $\mathbf{u} = (\mathbf{u}_\nu := v_\nu)_{\nu\in\mathcal{F}} \in \ell^2(\mathcal{F},V^*)$ for the vector of Legendre coefficients of the solution maps u. We finally introduce the operator $\mathbf{A}$ defined from $\ell^2(\mathcal{F},V)$ into $\ell^2(\mathcal{F},V^*) \simeq (\ell^2(\mathcal{F},V))^*$ by

$$\langle \mathbf{A}\mathbf{v},\mathbf{w}\rangle = \langle \mathcal{A}v,w\rangle, \qquad v,w\in\mathcal{V}_2. \tag{4.2.14}$$

where $\mathbf{v},\mathbf{w}\in\ell^2(\mathcal{F},V)$ the images of v,w by the isometry. The first duality product is defined by

$$\langle \mathbf{W},\mathbf{w}\rangle = \sum_{\nu\in\mathcal{F}} \langle \mathbf{W}_\nu,\mathbf{w}_\nu\rangle_{V^*,V}, \quad \mathbf{W}\in\ell^2(\mathcal{F},V^*), \quad \mathbf{w}\in\ell^2(\mathcal{F},V). \tag{4.2.15}$$

We associate to f considered as an element in $\mathcal{V}_2^* \simeq L^2(U,V^*,d\varrho)$, the vector $\mathbf{f}\in\ell^2(\mathcal{F},V^*)$. The system (4.2.11) is then equivalent to, the vector $\mathbf{u}$, repressing the solution map u, is solution of

$$\mathbf{A}\mathbf{u} = \mathbf{f} \quad \text{in} \quad \ell^2(\mathcal{F},V^*). \tag{4.2.16}$$

The operator $\mathbf{A}$ inherits the properties of the operator $\mathcal{A}$ and is then symmetric, bounded and definite positive. By linearity, $\mathbf{A}$ can also be seen as a bi-infinite symmetric matrix $(\mathbf{A}_{\nu\nu'})_{\nu,\nu'\in\mathcal{F}}$ of bounded, linear operators $\mathbf{A}_{\nu\nu'}\in\mathcal{L}(V,V^*)$ which are given as Riesz's representers

$$\langle \mathbf{A}_{\nu\nu'}w,w'\rangle_{V^*,V} = \mathcal{B}(w\otimes L_\nu, w'\otimes L_{\nu'}) \qquad w,w'\in V, \quad \nu,\nu'\in\mathcal{F}. \tag{4.2.17}$$

The infinite vector $\mathbf{f}$ can also be defined by its coordinates which satisfies

$$\langle \mathbf{f}_\nu,w\rangle_{V^*,V} = \mathcal{L}(w\otimes L_\nu), \qquad w\in V,\nu\in\mathcal{F}. \tag{4.2.18}$$

Let us remark that $\mathbf{f}_\nu = 0$ for any $\nu\neq 0_{\mathcal{F}}$. Indeed, for any such index, given $w\in V$

$$\mathcal{L}(w\otimes L_\nu) = \int_U \int_D f(x)w(x)L_\nu(y)dxd\varrho(y) = \int_U f(x)w(x)dx \int_D L_\nu(y)d\varrho(y) = 0, \tag{4.2.19}$$

because L_ν is orthogonal to $L_{0_{\mathcal{F}}}\equiv\mathbf{1}$ with respect to the measure ϱ. In the sequel, we only work with the system (4.2.16) and consider $\mathbf{A}$ as an infinite matrix of operators. Using the isometry between $\mathcal{V}_2$ and $\ell^2(\mathcal{F},V)$, we equip the latter with the norm $\|\cdot\|_E$ defined by

$$\|\mathbf{w}\|_E := \sqrt{\langle \mathbf{A}\mathbf{w},\mathbf{w}\rangle}, \quad \mathbf{w}\in\ell^2(\mathcal{F},V). \tag{4.2.20}$$

This notation is justified by $\|w\|_E = \|\mathbf{w}\|_E$ where $w\in\mathcal{V}_2$ and $\mathbf{w}$ its image by the isometry.

The primary objective of building adaptively Galerkin approximations u_Λ that are near optimal in the sense of (4.1.8) and (4.1.10) amount to building Galerkin projections $\mathbf{u}_\Lambda$ with respect to the formulation (4.2.16), that convergence to $\mathbf{u}$ the solution of the system (4.2.16) with the prescribed rates of (4.1.8) and (4.1.10).

Given a set of indices $\Lambda \subset \mathcal{F}$, we introduce the space

$$\ell^2(\Lambda, V) := \left\{ \mathbf{v} \in \ell^2(\mathcal{F}, V) : \mathbf{v}_\nu = 0 \text{ for } \nu \notin \Lambda \right\}. \tag{4.2.21}$$

The use of ℓ^2 in the previous notation is justified by, $\mathbf{v} \in \ell^2(\mathcal{F}, V)$ is supported in Λ implies $\sum_{\nu\in\Lambda} \|\mathbf{v}_\nu\|_V^2 = \|\mathbf{v}\|_{\ell^2(\mathcal{F},V)} < \infty$.

We introduce the notation $\mathrm{supp}(\mathbf{v})$ for the support of a vector $\mathbf{v} \in \ell^2(\mathcal{F}, V)$, so that if $\mathbf{v} \in \ell^2(\Lambda, V)$, we have necessarily $\mathrm{supp}(\mathbf{v}) \subset \Lambda$. We can now define the Galerkin projections associated with the system (4.2.16) as follows: $\mathbf{u}_\Lambda$ is the unique element in $\ell^2(\Lambda, V)$ such that

$$\langle \mathbf{A}\mathbf{u}_\Lambda, \mathbf{v}_\Lambda \rangle = \langle \mathbf{f}, \mathbf{v}_\Lambda \rangle, \quad \mathbf{v}_\Lambda \in \ell^2(\Lambda, V), \tag{4.2.22}$$

4.2.3 Properties of the operator A

Following the methodology of [30, 31, 49], we investigate in the present section the properties of the matrix $\mathbf{A}$, namely the decay of its entries $\mathbf{A}_{\nu,\nu'}$ and the norms that are related to it. We give in the next lemma, the explicit formulas of the operators $\mathbf{A}_{\nu,\nu'}$, and then examine the norm related to $\mathbf{A}$.

Lemma 4.2.1

For any $\nu, \nu' \in \mathcal{F}$ and any $w, w' \in V$, we have

$$\langle \mathbf{A}_{\nu\nu'} w, w' \rangle_{V^*,V} = \begin{cases} \displaystyle\int \bar{a} \nabla w \nabla w' & \text{if } \nu' = \nu, \\ \beta_{\nu_j} \displaystyle\int \psi_j \nabla w \nabla w' & \text{if } \nu' = \nu + e_j, \\ \beta_{\nu_j - 1} \displaystyle\int \psi_j \nabla w \nabla w' & \text{if } \nu' = \nu - e_j, \\ 0 & \text{otherwise,} \end{cases} \tag{4.2.23}$$

where the integrals are over D and $\beta_n = \frac{n+1}{\sqrt{2n+3}\sqrt{2n+1}}$ for any $n \geq 0$.

Proof: The univariate Legendre polynomials $(P_n)_{n\geq 0}$ satisfy the Bonnet recursion formula

$$P_0 = 1, \;\; P_1 = X \;\text{ and }\; (n+1)P_{n+1} = (2n+1)XP_n - nP_{n-1}, \quad n \geq 1,$$

As a consequence, the univariate Legendre polynomials $(L_n)_{n\geq 0}$ satisfy the recursion

$$L_0 = 1, \;\; L_1 = \sqrt{3}X \;\text{ and }\; \frac{n+1}{\sqrt{2n+3}} L_{n+1} = \frac{2n+1}{\sqrt{2n+1}} X L_n - \frac{n}{\sqrt{2n-1}} L_{n-1}, \;\; n \geq 1.$$

Therefore, one has

$$tL_n(t) = \beta_n L_{n+1}(t) + \beta_{n-1} L_{n-1}(t), \qquad n \geq 1, \quad t \in \Gamma,$$

where the β_j are defined as in the lemma. The orthonormality of the Legendre polynomials $(L_n)_{n\geq 0}$ with respect to the measure $\frac{dt}{2}$ combined with previous recursion yields that for any $(n,m) \in \mathbb{N}^2 - \{(0,0)\}$

$$\int_{-1}^{1} tL_n(t)L_m(t)\frac{dt}{2} = \begin{cases} \beta_n & \text{if } m = n+1, \\ \beta_{n-1} & \text{if } m = n-1, \\ 0 & \text{otherwise.} \end{cases}$$

Rematk that the identity also holds for $n = m = 0$. Therefore, the tensorized Legendre polynomials $(L_\nu)_{\nu\in\mathcal{F}}$ satisfies

$$\int_U y_j L_\nu(y)L_{\nu'}(y)d\varrho(y) = \int_{-1}^{1} tL_{\nu_j}(t)L_{\nu'_j}(t)\frac{dt}{2}\prod_{\substack{i\geq 1\\ i\neq j}}\int_{-1}^{1} L_{\nu_i}(t)L_{\nu'_i}(t)\frac{dt}{2} = \begin{cases} \beta_{\nu_j} & \text{if } \nu' = \nu + e_j, \\ \beta_{\nu_j-1} & \text{if } \nu' = \nu - e_j, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2.24}$$

Now given $w, w' \in V$, we have

$$\langle \mathbf{A}_{\nu\nu'} w, w'\rangle_{V^*,V} = \mathcal{B}(w\otimes L_\nu, w'\otimes L_{\nu'}) = \int_D \Big[\int_U a(x,y)L_\nu(y)L_{\nu'}(y)d\mu(y)\Big]\nabla w(x)\cdot\nabla w'(x)dx.$$

From the affine dependance of a on y, we infer

$$\int_U a(x,y)L_\nu(y)L_{\nu'}(y)d\varrho(y) = \bar{a}(x)\int_U L_\nu(y)L_{\nu'}(y)d\varrho(y) + \sum_{j\geq 1}\psi_j(x)\int_U y_j L_\nu(y)L_{\nu'}(y)d\varrho(y),$$

which in view of (4.2.24) implies

$$\int_U a(x,y)L_\nu(y)L_{\nu'}(y)d\mu(y) = \begin{cases} \bar{a}(x) & \text{if } \nu' = \nu, \\ \beta_{\nu_j}\psi_j(x) & \text{if } \nu' = \nu + e_j, \\ \beta_{\nu_j-1}\psi_j(x) & \text{if } \nu' = \nu - e_j, \\ 0 & \text{otherwise,} \end{cases}$$

The proof is then complete. ■

The previous sparsity result has an interesting implication on the inner product defined by $\mathbf{A}$. In the previous chapter, we introduced the notion of margin $\mathcal{M}(\Lambda)$ of a lower set Λ in Definition 3.2.3 as the indices $\nu \notin \Lambda$ such that $\nu - e_j \in \Lambda$ for some $j \geq 1$. For the purpose of the present analysis, we modify slightly the definition and generalize it to arbitrary subset $\mathcal{F}$.

Definition 4.2.2

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

$$\mathcal{M}(\Lambda) := \Big\{\nu \notin \Lambda \ ; \ \exists j > 0 \ : \nu - e_j \in \Lambda\Big\} \cup \Big\{\nu \notin \Lambda \ ; \ \exists j > 0 \ : \nu + e_j \in \Lambda\Big\} \quad (4.2.25)$$

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

This definition coincides with the definition (3.2.3) for a lower set Λ because for such sets if $\nu \notin \Lambda$, then necessarily for any j, $\nu + e_j \notin \Lambda$. Now given Λ arbitrary, $\nu \in \Lambda$ and $\nu' \notin \Lambda \cup \mathcal{M}(\Lambda)$, we have necessarily $\nu' \neq \nu$ and $\nu \neq \nu' \pm e_j$ for any $j \geq 1$. In view of Lemma 4.2.1, this implies

$$\mathbf{A}_{\nu\nu'} = 0, \qquad \nu \in \Lambda, \qquad \nu' \notin \Lambda \cup \mathcal{M}(\Lambda). \quad (4.2.26)$$

The previous identity implies in particular the following useful result.

Lemma 4.2.3

Let Λ be a non empty subset of $\mathcal{F}$, $\mathcal{M}$ the margin of Λ and $\mathbf{v}$, $\mathbf{w}$ two vector in $\ell^2(\mathcal{F}, V)$ such that $\mathbf{v}$ is supported in Λ and $\mathbf{w}$ is supported in $\mathcal{F} \setminus \{\Lambda \cup \mathcal{M}\}$, then

$$\langle \mathbf{A}\mathbf{v}, \mathbf{w} \rangle = 0. \quad (4.2.27)$$

The definition of the margin (3.2.3) in the previous chapter was motivated by the recursive relations relating Taylor coefficients t_ν to the coefficients $t_{\nu - e_j}$, which then is used to establish the energy reduction properties in Lemma 3.2.4. Here the modified definition is motivated by the sparsity of the matrix $\mathbf{A}$ which will also be used for residual energy reduction. Both definitions are consequences of the affine dependance of the diffusion coefficient a on the parameter y. In a more general case where a is a polynomials in y, the definition of the margin should be modified accordingly.

For the sake of our analysis, we need to give further properties of the matrix $\mathbf{A}$, or more precisely, the different norms related to $\mathbf{A}$. For notational convenience, as we have denoted by $\|\cdot\|_E$ both the norm over $\mathcal{V}_2$ defined in (4.2.3) and the norm over $\ell^2(\mathcal{F}, V)$ defined in (4.2.20), we simply denote $\|\cdot\|$ the norms of the Hilbert spaces $\mathcal{V}_2$ and $\ell^2(\mathcal{F}, V)$. Using the same notation every time is justified by the equality of each norm and its counterpart for vectors $(\mathbf{v}_\nu)_{\nu \in \mathcal{F}} \in \ell^2(\mathcal{F}, V)$ and functions $v = \sum_{\nu \in \mathcal{F}} \mathbf{v}_\nu L_\nu \in \mathcal{V}_2$. In particular, since the norms $\|\cdot\|$ and $\|\cdot\|_E$ are equivalent over $\mathcal{V}_2$ according to (4.2.4), then theirs counterpart over $\ell^2(\mathcal{F}, V)$ are also equivalent

$$\sqrt{r}\|\mathbf{v}\| \leq \|\mathbf{v}\|_E \leq \sqrt{R}\|\mathbf{v}\|, \quad \mathbf{v} \in \ell^2(\mathcal{F}, V). \quad (4.2.28)$$

In order to have a notation that is compatible with [30, 31, 49], we denote by $\|\cdot\|_S$ instead of $\|\cdot\|_{\ell^2(\mathcal{F}, V^*)}$ the norm over $\ell^2(\mathcal{F}, V^*)$. We have that $\ell^2(\mathcal{F}, V^*) \simeq (\ell^2(\mathcal{F}, V))^*$

and it can be easily checked by Cauchy-schwartz inequality that

$$\|\mathbf{g}\|_S = \sup_{\|\mathbf{v}\|=1} |\langle \mathbf{g}, \mathbf{v}\rangle|, \qquad \mathbf{g} \in \ell^2(\mathcal{F}, V^*). \tag{4.2.29}$$

where $\langle\cdot,\cdot\rangle$ is the duality pairing defined in (4.2.15). As for the two other norms, we also denote if needed by $\|\cdot\|_S$ the norm over $\mathcal{V}_2^*$. For operators defined from $\ell^2(\mathcal{F}, V)$ into $\ell^2(\mathcal{F}, V^*)$, such as $\mathbf{A}$, we use the same notation $\|\cdot\|_S$ to define the spectral norm

$$\|\mathbf{A}\|_S := \sup_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|_S = \sup_{\substack{\|\mathbf{v}\|=1\\ \|\mathbf{w}\|=1}} |\langle \mathbf{A}\mathbf{v}, \mathbf{w}\rangle|. \tag{4.2.30}$$

Similarily, for operator defined from $\ell^2(\mathcal{F}, V^*)$ into $\ell^2(\mathcal{F}, V)$, such as $\mathbf{A}^{-1}$, we define the following norm

$$\|\mathbf{A}^{-1}\|_S = \sup_{\substack{\mathbf{g}\in\ell^2(\mathcal{F},V^*)\\ \|\mathbf{g}\|_S=1}} \|\mathbf{A}^{-1}\mathbf{g}\|. \tag{4.2.31}$$

The following lemmas gives results on the different equivalencies involving the defined norms

Lemma 4.2.4

For any $\mathbf{v} \in \ell^2(\mathcal{F}; V)$ and $\mathbf{g} \in \ell^2(\mathcal{F}; V^*)$, it holds

$$r\|\mathbf{v}\| \le \|\mathbf{A}\mathbf{v}\|_S \le R\|\mathbf{v}\|, \tag{4.2.32}$$

$$R^{-1}\|\mathbf{g}\|_S \le \|\mathbf{A}^{-1}\mathbf{g}\| \le r^{-1}\|\mathbf{g}\|_S. \tag{4.2.33}$$

Consequently, the condition number $\kappa(\mathbf{A}) := \|\mathbf{A}\|_S\|\mathbf{A}^{-1}\|_S$ of $\mathbf{A}$ satisfies

$$\kappa(\mathbf{A}) \le \frac{R}{r}. \tag{4.2.34}$$

Proof: Let $\mathbf{v} \in \ell^2(\mathcal{F}, V)$. On the one hand, the assumption $\mathbf{UEA}(r, R)$ implies

$$r\|\mathbf{v}\|^2 \le |\langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle| \le \|\mathbf{A}\mathbf{v}\|_S\|\mathbf{v}\|,$$

which gives the first inequality in 4.2.32. On the other hand, by Cauchy Schwartz formula and (4.2.28), we have for any $\mathbf{w} \in \ell^2(\mathcal{F}, V)$

$$|\langle \mathbf{A}\mathbf{v}, \mathbf{w}\rangle| \le |\langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle| \cdot |\langle \mathbf{A}\mathbf{w}, \mathbf{w}\rangle| \le R\|\mathbf{v}\|\|\mathbf{w}\|,$$

which implies, using the definition (4.2.29), the second inequality in (4.2.32). The inequalities (4.2.33) and (4.2.34) are straightforward applications of (4.2.32) with $\mathbf{A}^{-1}\mathbf{g}$ instead of $\mathbf{v}$. ■

Lemma 4.2.5

For any $\mathbf{v} \in \ell^2(\mathcal{F}; V)$,

$$\sqrt{r}\|\mathbf{v}\|_E \leq \|\mathbf{A}^{-1}\|_S^{-1/2}\|\mathbf{v}\|_E \leq \|\mathbf{A}\mathbf{v}\|_S \leq \|\mathbf{A}\|_S^{1/2}\|\mathbf{v}\|_E \leq \sqrt{R}\|\mathbf{v}\|_E \tag{4.2.35}$$

Proof: Let us first remark that

$$\|\mathbf{v}\|_E \leq \sqrt{\|\mathbf{A}\|_S}\|\mathbf{v}\|, \qquad \mathbf{v} \in \ell^2(\mathcal{F}, V). \tag{4.2.36}$$

Indeed, for $\mathbf{v} \in \ell^2(\mathcal{F}, V)$, we have $\|\mathbf{v}\|_E^2 = |\langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle| \leq \|\mathbf{A}\mathbf{v}\|_S\|\mathbf{v}\| \leq \|\mathbf{A}\|_S\|\mathbf{v}\|^2$. Now, the Cauchy-Schwartz formula applied with the scalar product $\langle \mathbf{A}\cdot, \cdot\rangle$ yields

$$|\langle \mathbf{A}\mathbf{v}, \mathbf{w}\rangle| \leq \|\mathbf{v}\|_E\|\mathbf{w}\|_E, \quad \mathbf{v}, \mathbf{w} \in \ell^2(\mathcal{F}, V).$$

This combined with the inequalities (4.2.36) and (4.2.32) implies

$$|\langle \mathbf{A}\mathbf{v}, \mathbf{w}\rangle| \leq \|\mathbf{v}\|_E\sqrt{\|\mathbf{A}\|_S}\|\mathbf{w}\| \leq \|\mathbf{v}\|_E\sqrt{R}\|\mathbf{w}\|,$$

which implies the two last inequalities in (4.2.35). As for the two first inequalities, since $\mathbf{A}$ is a isomorphism from $\ell^2(\mathcal{F}, V)$ into $\ell^2(\mathcal{F}, V^*)$, then proving them is equivalent to prove that for any $\mathbf{g} \in \ell^2(\mathcal{F}, V^*)$

$$\sqrt{r}\|\mathbf{A}^{-1}\mathbf{g}\|_E \leq \|\mathbf{A}^{-1}\|_S^{-1/2}\|\mathbf{A}^{-1}\mathbf{g}\|_E \leq \|\mathbf{g}\|_S,$$

or equivalently

$$\|\mathbf{A}^{-1}\mathbf{g}\|_E \leq \sqrt{\|\mathbf{A}^{-1}\|_S}\|\mathbf{g}\|_S \leq \sqrt{r^{-1}}\|\mathbf{g}\|_S.$$

The second part in the last inequality is a straightforward application of (4.2.33). Now, giving $\mathbf{g} \in \ell^2(\mathcal{F}, V^*)$, we have by the definition of $\|\cdot\|_E$

$$\|\mathbf{A}^{-1}\mathbf{g}\|_E^2 = |\langle \mathbf{g}, \mathbf{A}^{-1}\mathbf{g}\rangle| \leq \|\mathbf{g}\|_S\|\mathbf{A}^{-1}\mathbf{g}\| \leq \|\mathbf{g}\|_S\|\mathbf{A}^{-1}\|_S\|\mathbf{g}\|_S,$$

which implies the first part in (4.2.3). ■

Having established these interesting properties of the operator $\mathbf{A}$ and related norms, we have now the necessary tools for adjusting the techniques of adaptive wavelet methods studied in [30, 31, 49] to the present setting. To start, given Λ an index set and $\mathbf{u}_\Lambda$ the Galerkin projection in $\ell^2(\Lambda, V)$ as in (4.2.22), we have by the previous inequalities

$$\|\mathbf{u} - \mathbf{u}_\Lambda\| \leq \frac{1}{r}\|\mathbf{A}\mathbf{u} - \mathbf{A}\mathbf{u}_\Lambda\|_S = \frac{1}{r}\|\mathbf{f} - \mathbf{A}\mathbf{u}_\Lambda\|_S. \tag{4.2.37}$$

Therefore, searching for near optimality with Galerkin projections amounts to find index sets $(\Lambda_k)_{k\geq 1}$ that yields a convergence for $\|\mathbf{f} - \mathbf{A}\mathbf{u}_{\Lambda_k}\|_S$ with rate $(n+1)^{-s^*}$ where $n = n(k) = \#(\Lambda_k)$.

4.3 Reduction of Galerkin residuals

Our strategy for building the previously described sets is inspired from the adaptive algorithms in [30, 31, 49] involving residuals. For Λ a subset of $\mathcal{F}$, we define the residual associated to Λ by

$$\mathbf{r}_\Lambda := \mathbf{A}\mathbf{u} - \mathbf{A}\mathbf{u}_\Lambda = \mathbf{f} - \mathbf{A}\mathbf{u}_\Lambda \in \ell^2(\mathcal{F}, V^*). \tag{4.3.1}$$

Given an index set $\mathcal{R}$ in $\mathcal{F}$ and $\mathbf{v}$ a vector in $\ell^2(\mathcal{F}, V)$, we denote by $\mathbf{P}_\mathcal{R}\mathbf{v}$ the orthogonal projection with respect to the norm $\|\cdot\|$ of $\mathbf{v}$ into the space $\ell^2(\mathcal{R}, V)$. The projection $\mathbf{P}_\mathcal{R}\mathbf{v}$ is simply the vector in $\ell^2(\mathcal{F}, V)$ that agrees with $\mathbf{v}$ in the coordinates corresponding to indices $\nu \in \mathcal{R}$ and have null coordinates in $\nu \notin \mathcal{R}$. We define similarly $\mathbf{P}_\mathcal{R}\mathbf{g}$ if $\mathbf{g} \in \ell^2(\mathcal{F}, V^*)$. We recall that the Galerkin solution $\mathbf{u}_\Lambda$ is the unique solution in $\ell^2(\Lambda, \mathcal{F})$ of the problem

$$\mathbf{P}_\Lambda \mathbf{r}_\Lambda = 0, \quad \text{or equivalently} \quad \mathbf{P}_\Lambda \mathbf{A}\mathbf{u}_\Lambda = \mathbf{P}_\Lambda \mathbf{A}\mathbf{u}. \tag{4.3.2}$$

We shall use the notation $\mathbf{u}_{\Lambda,\nu} := (\mathbf{u}_\Lambda)_\nu$ to denote the coordinate of the vector $\mathbf{u}_\Lambda \in \ell^2(\Lambda, V)$. Our first result is concerned with the quantification of the norms of the residuals $\mathbf{r}_\Lambda$ for arbitrary subsets Λ of $\mathcal{F}$.

Lemma 4.3.1

Let $\Lambda \subset \mathcal{F}$ be a non empty index set containing $0_\mathcal{F}$, $\mathcal{M}$ the margin of Λ and $\mathbf{u}_\Lambda$ the Galerkin approximation of $\mathbf{u}$ associated with Λ. One has $\mathbf{r}_\Lambda \in \ell^2(\mathcal{M}, V^*)$ and

$$\|\mathbf{r}_\Lambda\|_S = \Big| \sum_{\nu \in \mathcal{M}} \|w_{\Lambda,\nu}\|_V^2 \Big|^{\frac{1}{2}}, \tag{4.3.3}$$

where the functions $w_{\Lambda,\nu}$ are the solutions in V of the systems

$$-\Delta w = \ div\ \phi_\nu \ \text{ in } \ D, \quad w = 0 \ \text{ on } \ \partial D, \tag{4.3.4}$$

with for each $\nu \in \mathcal{M}$

$$\phi_\nu = \phi_\nu(\Lambda) := \sum_{\substack{j \geq 1 \\ \nu + e_j \in \Lambda}} \beta_{\nu_j} \psi_j \nabla \mathbf{u}_{\Lambda, \nu + e_j} + \sum_{\substack{j \geq 1 \\ \nu - e_j \in \Lambda}} \beta_{\nu_j - 1} \psi_j \nabla \mathbf{u}_{\Lambda, \nu - e_j}, \tag{4.3.5}$$

where the sequence $(\beta_n)_{n \geq 0}$ is as in Lemma 4.2.1.

Proof: The indices sets Λ, $\mathcal{M}$ and $\mathcal{Q} := \mathcal{F} \setminus \{\Lambda \cup \mathcal{M}\}$ form a disjoint union of $\mathcal{F}$, therefore given $\mathbf{v} \in \ell^2(\mathcal{F}, V)$, one has

$$\langle \mathbf{r}_\Lambda, \mathbf{v} \rangle = \langle \mathbf{r}_\Lambda, \mathbf{P}_\Lambda \mathbf{v} + \mathbf{P}_\mathcal{M} \mathbf{v} + \mathbf{P}_\mathcal{Q} \mathbf{v} \rangle = \langle \mathbf{r}_\Lambda, \mathbf{P}_\Lambda \mathbf{v} \rangle + \langle \mathbf{r}_\Lambda, \mathbf{P}_\mathcal{M} \mathbf{v} \rangle + \langle \mathbf{r}_\Lambda, \mathbf{P}_\mathcal{Q} \mathbf{v} \rangle.$$

The Galerkin equation (4.3.2) implies $\langle \mathbf{r}_\Lambda, \mathbf{P}_\Lambda \mathbf{v}\rangle = \langle \mathbf{P}_\Lambda \mathbf{r}_\Lambda, \mathbf{P}_\Lambda \mathbf{v}\rangle = 0$. Now, from the definition (4.3.1) of residuals $\langle \mathbf{r}_\Lambda, \mathbf{P}_\mathcal{Q}\mathbf{v}\rangle = \langle \mathbf{f}, \mathbf{P}_\mathcal{Q}\mathbf{v}\rangle - \langle \mathbf{A}\mathbf{u}_\Lambda, \mathbf{P}_\mathcal{Q}\mathbf{v}\rangle = 0$, where we have used that $\mathbf{f}$ is supported in $0_\mathcal{F}$, $0_\mathcal{F} \notin \mathcal{Q}$ and Lemma 4.2.3. We infer that

$$\langle \mathbf{r}_\Lambda, \mathbf{v}\rangle = \langle \mathbf{r}_\Lambda, \mathbf{P}_\mathcal{M}\mathbf{v}\rangle = \langle \mathbf{f}, \mathbf{P}_\mathcal{M}\mathbf{v}\rangle - \langle \mathbf{A}\mathbf{u}_\Lambda, \mathbf{P}_\mathcal{M}\mathbf{v}\rangle = -\langle \mathbf{A}\mathbf{u}_\Lambda, \mathbf{P}_\mathcal{M}\mathbf{v}\rangle,$$

where we have used again that $\mathbf{f}$ is supported in $0_\mathcal{F}$ and $0_\mathcal{F} \notin \mathcal{M}$. This show that $\mathbf{r}_\Lambda$ is supported in $\mathcal{M}$ and it implies

$$\langle \mathbf{r}_\Lambda, \mathbf{v}\rangle = -\sum_{\nu\in\mathcal{M}} \sum_{\nu'\in\Lambda} \langle \mathbf{A}_{\nu\nu'}\mathbf{u}_{\Lambda,\nu'}, \mathbf{v}_\nu\rangle_{V^*,V}.$$

Since $\mathbf{A}_{\nu\nu'} = 0$ unless $\nu = \nu'$ or $\nu = \nu' \pm e_j$ for some j and using that $\Lambda \cap \mathcal{M} = \emptyset$, we have

$$\langle \mathbf{r}_\Lambda, \mathbf{v}\rangle \quad = -\sum_{\nu\in\mathcal{M}} \sum_{\substack{j\geq 1\\ \nu+e_j\in\Lambda}} \langle \mathbf{A}_{\nu,\nu+e_j}\mathbf{u}_{\Lambda,\nu+e_j}, \mathbf{v}_\nu\rangle_{V^*,V} - \sum_{\nu\in\mathcal{M}} \sum_{\substack{j\geq 1\\ \nu-e_j\in\Lambda}} \langle \mathbf{A}_{\nu,\nu-e_j}\mathbf{u}_{\Lambda,\nu-e_j}, \mathbf{v}_\nu\rangle_{V^*,V}.$$

Using the explicit formulas of the operators $\mathbf{A}_{\nu\nu'}$ given in Lemma (4.2.1), we get

$$\langle \mathbf{r}_\Lambda, \mathbf{v}\rangle = -\sum_{\nu\in\mathcal{M}} \int_D \phi_\nu(\Lambda)\nabla \mathbf{v}_\nu,$$

where the functions $\phi_\nu(\Lambda)$ are as in (4.3.5). The coordinate operators $\mathbf{r}_{\Lambda,\nu} \in V^*$ for $\nu \in \mathcal{M}$ act individually on the corresponding coordinates $\mathbf{v}_\nu$ of $\mathbf{v}$ for $\nu \in \mathcal{M}$ and are defined by

$$\langle \mathbf{r}_{\Lambda,\nu}, w\rangle_{V^*,V} := -\int_D \phi_\nu(\Lambda)\nabla w, \quad w \in V.$$

From the definition of the norm $\|\cdot\|_S$, which is equal to $\|\cdot\|_{\ell^2(\mathcal{F},V)}$, we deduce that the norm of $\mathbf{r}_\Lambda$ can be obtained from the norm of its coordinates $\mathbf{r}_{\Lambda,\nu}$ according to

$$\|\mathbf{r}_\Lambda\|_S = \Big|\sum_{\nu\in\mathcal{M}} \|\mathbf{r}_{\Lambda,\nu}\|_{V^*}^2\Big|^{\frac{1}{2}} \quad \text{with} \quad \|\mathbf{r}_{\Lambda,\nu}\|_S = \sup_{\|w\|_V=1} \Big|\int_D \phi_\nu(\Lambda)\nabla w\Big|. \tag{4.3.6}$$

It can be checked easily that the previous supremum are attained on $\frac{w_{\Lambda,\nu}}{\|w_{\Lambda,\nu}\|_V} \in V$, where $w_{\Lambda,\nu}$ is the unique solution in V of the PDE (4.3.4). Substituting by $\frac{w_{\Lambda,\nu}}{\|w_{\Lambda,\nu}\|_V}$ and using Green identity, we deduce

$$\|\mathbf{r}_{\Lambda,\nu}\|_S = \frac{1}{\|w_{\Lambda,\nu}\|_V}\Big|\int_D \phi_\nu(\Lambda)\nabla w_{\Lambda,\nu}\Big| = \frac{1}{\|w_{\Lambda,\nu}\|_V}\Big|\int_D \operatorname{div}\, \phi_\nu(\Lambda)w_{\Lambda,\nu}\Big| = \frac{1}{\|w_{\Lambda,\nu}\|_V}\Big|\int_D \Delta w_{\Lambda,\nu} w_{\Lambda,\nu}\Big|,$$

so that that using again Green identity, we obtain $\|\mathbf{r}_{\Lambda,\nu}\|_S = \|w_{\Lambda,\nu}\|_V$. This completes the proof. ∎

Before continuing our analysis, we should make the following remark

Remark 4.3.2

The assumption $0_{\mathcal{F}} \in \Lambda$ in the previous lemma is redundant if Λ is lower. Moreover, for such sets, $\nu + e_j \notin \Lambda$ for any $\nu \in \mathcal{M}(\Lambda)$ and $j \geq 0$, therefore the functions $\phi_\nu(\Lambda)$ are given by

$$\phi_\nu = \phi_\nu(\Lambda) := \sum_{\substack{j\geq 1 \\ \nu - e_j \in \Lambda}} \beta_{\nu_j - 1} \psi_j \nabla \mathbf{u}_{\Lambda, \nu - e_j}. \tag{4.3.7}$$

The previous lemma and its proof show us that the residual $\mathbf{r}_\Lambda$ is supported in $\mathcal{M} := \mathcal{M}(\Lambda)$ and that the norms of its coordinates are given by $\|\mathbf{r}_{\Lambda,\nu}\|_{V^*} = \|w_{\Lambda,\nu}\|_V$, therefore for any set of indices $\tilde{\Lambda}$, the vector $\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}_\Lambda$ is supported in $\mathcal{M} \cap \tilde{\Lambda}$ and its norm is given by

$$\|\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}_\Lambda\|_S = \|\mathbf{P}_{\tilde{\Lambda}\cap\mathcal{M}}\mathbf{r}_\Lambda\|_S = \Big| \sum_{\nu\in\tilde{\Lambda}\cap\mathcal{M}} \|w_{\Lambda,\nu}\|_V^2 \Big|^{\frac{1}{2}}, \tag{4.3.8}$$

where the functions $w_{\Lambda,\nu}$ are as in Lemma 4.3.1. According to [30, 31], given an index set Λ, an easy way to build an index set $\tilde{\Lambda}$ containing Λ with the reduction propriety (4.3.11) below is by choosing $\tilde{\Lambda}$ satisfying the property

$$\|\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}_\Lambda\|_S \geq \theta \|\mathbf{r}_\Lambda\|_S \tag{4.3.9}$$

where θ is any number in $]0, 1[$. In view of the localization of the residual property (4.3.8), It is then obvious that instead of searching for an arbitrary set $\tilde{\Lambda}$ with the bulk property (4.3.9), it is sufficient to search for it with the property that it contains Λ and it is contained in $\Lambda \cup \mathcal{M}$. Let us recall that this also was the case when we worked with Taylor residuals in Chapter 3. The following lemma provides further results in this direction.

Lemma 4.3.3

Let $\Lambda \subset \mathcal{F}$ be a set containing $0_{\mathcal{F}}$ and $\mathbf{r}_\Lambda$ the residual associated with Λ. If $0 < \theta < \min(1, \sqrt{\kappa(\mathbf{A})})$ and $\mathcal{S}$ is any subset of $\mathcal{M}(\Lambda)$ such that

$$\Big(\sum_{\nu\in\mathcal{S}} \|w_{\Lambda,\nu}\|^2\Big)^{\frac{1}{2}} \geq \theta \Big(\sum_{\nu\in\mathcal{M}} \|w_{\Lambda,\nu}\|^2\Big)^{\frac{1}{2}}, \tag{4.3.10}$$

then $\tilde{\Lambda} = \Lambda \cup \mathcal{S}$ contains Λ and satisfies

$$\|\mathbf{u} - \mathbf{u}_{\tilde{\Lambda}}\|_E \leq \delta \|\mathbf{u} - \mathbf{u}_\Lambda\|_E, \tag{4.3.11}$$

where $\delta = \sqrt{1 - \frac{\theta^2}{\kappa(\mathbf{A})}}$

Proof: The formulas (4.3.3) and (4.3.8) translates the bulk inequality (4.3.10) into $\|\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}_\Lambda\|_S \geq \theta\|\mathbf{r}_\Lambda\|_S$. In addition, we have

$$\|\mathbf{A}(\mathbf{u}_{\tilde{\Lambda}} - \mathbf{u}_\Lambda)\|_S = \|\mathbf{r}_{\tilde{\Lambda}} - \mathbf{r}_\Lambda\|_S \geq \|\mathbf{P}_{\tilde{\Lambda}}(\mathbf{r}_{\tilde{\Lambda}} - \mathbf{r}_\Lambda)\|_S = \|\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}_\Lambda\|_S$$

where we have used the Galerkin identity $\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}_{\tilde{\Lambda}} = 0$. Combining the two last inequalities and the inequality (4.2.35), we obtain

$$\theta\|\mathbf{A}^{-1}\|^{-\frac{1}{2}}\|\mathbf{u}-\mathbf{u}_\Lambda\|_E \le \theta\|\mathbf{A}(\mathbf{u}-\mathbf{u}_\Lambda)\|_S = \theta\|\mathbf{r}_\Lambda\|_S \le \|\mathbf{P}_{\tilde{\Lambda}}\mathbf{r}_\Lambda\|_S \le \|\mathbf{A}(\mathbf{u}_{\tilde{\Lambda}}-\mathbf{u}_\Lambda)\|_S \le \|\mathbf{A}\|^{\frac{1}{2}}\|\mathbf{u}_{\tilde{\Lambda}}-\mathbf{u}_\Lambda\|_E,$$

therefore

$$\theta\kappa(\mathbf{A})^{-\frac{1}{2}}\|\mathbf{u}-\mathbf{u}_\Lambda\|_E \le \|\mathbf{u}_{\tilde{\Lambda}}-\mathbf{u}_\Lambda\|_E.$$

By the orthogonality of the Galerkin solutions with respect to the energy norm $\|\cdot\|_E$, we have

$$\|\mathbf{u}-\mathbf{u}_\Lambda\|_E^2 = \|\mathbf{u}-\mathbf{u}_{\tilde{\Lambda}}\|_E^2 + \|\mathbf{u}_{\tilde{\Lambda}}-\mathbf{u}_\Lambda\|_E^2,$$

therefore

$$\|\mathbf{u}-\mathbf{u}_{\tilde{\Lambda}}\|_E \le \sqrt{1-\frac{\theta^2}{\kappa(\mathbf{A})}}\|\mathbf{u}-\mathbf{u}_\Lambda\|_E,$$

which finishes the proof. ■

Our next result is concerned with the control of the cardinality of the set $\tilde{\Lambda}$ under certain assumptions on θ and $\mathcal{S}$ and is given in the next lemma. We use in particular the arguments of [49] which shows that a control of the cardinality can be obtained if θ is small enough. In order to lighten our notation, we introduce the notation $(c_\nu)_{\nu\in\mathcal{F}}$ for the sequence of the V-norm of the Legendre coefficients v_ν of u defined in (4.1.7), i.e.

$$c_\nu := \|v_\nu\|_V, \qquad \nu \in \mathcal{F}. \tag{4.3.12}$$

We recall that, according to Theorem (4.1.1), this sequence belongs to $\ell^p_m(\mathcal{F})$ and any sequence $(\Lambda_n^L)_{n\ge1}$ of nested sets with $\#(\Lambda_n^L) = n$ and Λ_n corresponds to n largest value of c_ν can be used in the best n-term approximation (4.1.8). In view of (4.2.9), this sequence of index sets also yields convergence of Galerkin approximations with convergence rates (4.1.8).

Lemma 4.3.4

In the previous lemma, if in addition $0 < \theta < \kappa(\mathbf{A})^{-\frac{1}{2}}$ and $\mathcal{S}$ is the smallest set in $\mathcal{M}(\Lambda)$ satisfying (4.3.10), then

$$\#(\mathcal{S}) \le C_\theta\|(c_\nu)\|_{\ell_p}^{1/s^*}\|\mathbf{r}_\Lambda\|^{-1/s^*}, \qquad s^* = \frac{1}{p}-\frac{1}{2}. \tag{4.3.13}$$

where C_θ is a constant.

Proof: In view of (4.1.8) and (4.2.36), we have

$$\|\mathbf{u}-\mathbf{P}_{\Lambda_n^L}\mathbf{u}\|_E \le \sqrt{\|\mathbf{A}\|_S}\|\mathbf{u}-\mathbf{P}_{\Lambda_n^L}\mathbf{u}\| \le \sqrt{\|\mathbf{A}\|_S}\|(c_\nu)\|_{\ell^p(\mathcal{F})}(n+1)^{-s^*},$$

hence $\|\mathbf{u}-\mathbf{P}_{\Lambda_n^L}\mathbf{u}\|_E$ tends to 0 as n tends to ∞. Now, giving that $0<\theta<\kappa(\mathbf{A})^{-\frac{1}{2}}$, we fix a number $\lambda>0$ such that $\theta\leq\kappa(\mathbf{A})^{-\frac{1}{2}}(1-\lambda^2)^{\frac{1}{2}}$ and let n be the smallest integer such that

$$\|\mathbf{u}-\mathbf{P}_{\Lambda_n^L}\mathbf{u}\|_E\leq\lambda\|\mathbf{u}-\mathbf{u}_\Lambda\|_E.$$

On the one hand by Lemma 4.2.5

$$\lambda\|\mathbf{A}\|_S^{-\frac{1}{2}}\|\mathbf{r}_\Lambda\|_S=\lambda\|\mathbf{A}\|_S^{-\frac{1}{2}}\|\mathbf{A}(\mathbf{u}-\mathbf{u}_\Lambda)\|_S\leq\lambda\|\mathbf{u}-\mathbf{u}_\Lambda\|_E.$$

On the other hand, by minimality of n

$$\lambda\|\mathbf{u}-\mathbf{u}_\Lambda\|_E\leq\|\mathbf{u}-\mathbf{P}_{\Lambda_{n-1}^L}\mathbf{u}\|_E\leq\sqrt{\|\mathbf{A}\|_S}\|(c_\nu)\|_{\ell^p(\mathcal{F})}n^{-s^*}.$$

Combining the two inequalities, we deduce

$$n\leq C_\theta\|(c_\nu)\|_{\ell^p(\mathcal{F})}^{1/s^*}\|\mathbf{r}_\Lambda\|_S^{-1/s^*}$$

where $C_\theta:=\|\mathbf{A}\|_S^{1/s^*}\lambda^{-1/s^*}$ depends only on θ and $\|\mathbf{A}\|_S$. At this stage, we observe that proving $\#(\mathcal{S})\leq n$ finishes the proof. Let $\hat{\Lambda}=\Lambda\cup\Lambda_n^L$. The optimality of the Galerkin projection $\mathbf{u}_{\hat{\Lambda}}$ yields

$$\|\mathbf{u}-\mathbf{u}_{\hat{\Lambda}}\|_E\leq\|\mathbf{u}-\mathbf{P}_{\Lambda_n^L}\mathbf{u}\|_E\leq\lambda\|\mathbf{u}-\mathbf{u}_\Lambda\|_E.$$

Now by Galerkin orthogonality, we have

$$\|\mathbf{u}-\mathbf{u}_\Lambda\|_E^2=\|\mathbf{u}-\mathbf{u}_{\hat{\Lambda}}\|_E^2+\|\mathbf{u}_{\hat{\Lambda}}-\mathbf{u}_\Lambda\|_E^2.$$

Combining the two previous inequalities using how θ and λ are related, we obtain

$$\|\mathbf{u}_{\hat{\Lambda}}-\mathbf{u}_\Lambda\|_E^2\geq(1-\lambda^2)\|\mathbf{u}-\mathbf{u}_\Lambda\|_E^2\geq\theta^2\kappa(\mathbf{A})\|\mathbf{u}-\mathbf{u}_\Lambda\|_E^2.$$

Using the inequalities of Lemma 4.2.5, we deduce

$$\begin{aligned}\|\mathbf{P}_{\hat{\Lambda}}(\mathbf{A}\mathbf{u}-\mathbf{A}\mathbf{u}_\Lambda)\|_S&=\|\mathbf{P}_{\hat{\Lambda}}(\mathbf{A}\mathbf{u}_{\hat{\Lambda}}-\mathbf{A}\mathbf{u}_\Lambda)\|_S\\&=\|\mathbf{P}_{\hat{\Lambda}}\mathbf{A}(\mathbf{u}_{\hat{\Lambda}}-\mathbf{u}_\Lambda)\|_S\\&\geq\|\mathbf{A}^{-1}\|_S^{-\frac{1}{2}}\|\mathbf{u}_{\hat{\Lambda}}-\mathbf{u}_\Lambda\|_E\\&\geq\|\mathbf{A}^{-1}\|_S^{-\frac{1}{2}}\theta\sqrt{\kappa(\mathbf{A})}\|\mathbf{u}-\mathbf{u}_\Lambda\|_E\\&=\theta\sqrt{\|\mathbf{A}\|_S}\|\mathbf{u}-\mathbf{u}_\Lambda\|_E\\&\geq\theta\|\mathbf{A}\mathbf{u}-\mathbf{A}\mathbf{u}_\Lambda\|_S,\end{aligned}$$

where we have used that $\mathbf{A}$ is a symmetric positive definite operator in the third line. The previous result can be written equivalently as

$$\|\mathbf{P}_{\hat{\Lambda}}\mathbf{r}_\Lambda\|_S\geq\theta\|\mathbf{r}_\Lambda\|_S.$$

Since by the observation (4.3.8), we have $\mathbf{P}_{\hat{\Lambda}}\mathbf{r}_\Lambda=\mathbf{P}_{\{\hat{\Lambda}\cap\mathcal{M}(\Lambda)\}}\mathbf{r}_\Lambda$, then $\mathbf{P}_{\hat{\Lambda}}\mathbf{r}_\Lambda=\mathbf{P}_{\{\Lambda_k^L\cap\mathcal{M}(\Lambda)\}}\mathbf{r}_\Lambda$, so that

$$\|\mathbf{P}_{\{\Lambda_k^L\cap\mathcal{M}(\Lambda)\}}\mathbf{r}_\Lambda\|_S\geq\theta\|\mathbf{r}_\Lambda\|_S.$$

The definition of the set $\mathcal{S}$ as the smallest subset of $\mathcal{M}$ with the bulk property implies

$$\#(\mathcal{S}) \leq \#(\Lambda_n^L \cap \mathcal{M}(\Lambda)) \leq \#(\Lambda_n^L) = n, \tag{4.3.14}$$

and completes the proof. ■

In the previous proof, we have only used that $(c_\nu)_{\nu\in\mathcal{F}}$ is ℓ^p-summable. However, the sequence $(c_\nu)_{\nu\in\mathcal{F}}$ is ℓ^p_m-summable. This stronger summability allows us to prove that the previous lemma can also holds for lower sets. More precisely, we have the following. We recall that given Λ a lower set, the definition of a set $\mathcal{S}$ being lower in lower in $\mathcal{M}(\Lambda)$ is given in (3.2.20). In particular, it should be noted, see (3.2.22), that in such case $\Lambda' = \Lambda \cup \mathcal{S}$ is a lower set.

Lemma 4.3.5

If in Lemma (4.3.3) we have in addition $0 < \theta < \kappa(\mathbf{A})^{-\frac{1}{2}}$ and Λ is a non empty lower set. Then with $\mathcal{M}$ the margin of Λ and $\mathcal{S}$ is the smallest lower set in $\mathcal{M}$ satisfying (4.3.10), we have

$$\#(\mathcal{S}) \leq C_\theta \|(c_\nu)\|^{1/s^*}_{\ell^m_p(\mathcal{F})} \|\mathbf{r}_\Lambda\|^{-1/s^*}, \tag{4.3.15}$$

with the same constant C_θ used in Lemma 4.3.4.

Proof: The proof is similar to the proof of Lemma 4.3.4 but uses the $\ell^p_m(\mathcal{F})$ summability of the sequence $(c_\nu)_{\nu\in\mathcal{F}}$. According to (4.1.10), we have that for any $n \geq 1$, the index set Λ_n^{L*} is lower and

$$\|\mathbf{u}-\mathbf{P}_{\Lambda_n^{L*}}\mathbf{u}\|_E \leq \sqrt{\|\mathbf{A}\|_S}\|\mathbf{u}-\mathbf{P}_{\Lambda_n^{L*}}\mathbf{u}\| \leq \|(c_\nu)\|_{\ell^p_m(\mathcal{F})}(n+1)^{-s^*}, \quad s^* = \frac{1}{p}-\frac{1}{2}. \tag{4.3.16}$$

We consider λ and n as in the proof of Lemma 4.3.4 but n here depends on the lower set Λ_n^{L*}. By the same arguments there, we obtain

$$n \leq C_\theta \|(c_\nu)\|^{1/s}_{\ell^p_m(\mathcal{F})} \|\mathbf{r}_\Lambda\|^{-1/s}_S \tag{4.3.17}$$

where $C_\theta := \|\mathbf{A}\|_S^{\frac{1}{s}} \lambda^{-\frac{1}{s}}$ is the same constant. To finish the proof, we only need to show that $\#(\mathcal{S}) \leq n$. To this end, we consider $\hat{\Lambda} = \Lambda \cup \Lambda_k^{L*}$. By the same arguments of the proof of Lemma 4.3.5 , we can prove

$$\|\mathbf{P}_{\hat{\Lambda}}\mathbf{r}_\Lambda\|_S \geq \theta \|\mathbf{r}_\Lambda\|_S.$$

therefore $\|\mathbf{P}_{\{\hat{\Lambda}\cap\mathcal{M}(\Lambda)\}}\mathbf{r}_\Lambda\|_S \geq \theta\|\mathbf{r}_\Lambda\|_S$. Since $\hat{\Lambda}$ is the union of two lower sets, then it is lower, which implies that $\hat{\Lambda} \cap \mathcal{M}(\Lambda)$ is lower in $\mathcal{M}(\Lambda)$. By the minimality of the cardinality of $\mathcal{S}$, we deduce that

$$\#(\mathcal{S}) \leq \#(\hat{\Lambda} \cap \mathcal{M}) = \#(\Lambda_n^{L*} \cap \mathcal{M}) \leq \#(\Lambda_n^{L*}) = n, \tag{4.3.18}$$

which complete the proof. ■

Similarly to chapter 3, given an index set Λ, which is here not necessarily lower, we are able to enrich Λ by a set $\mathcal{S} \subset \mathcal{M}(\Lambda)$, with a controlled cardinality as in (4.3.13) and (4.3.15), that yields a reduction on the approximation error $\|\mathbf{u} - \mathbf{u}_\Lambda\|$. This can then be a starting point for a bulk chasing procedure.

4.4 Bulk chasing algorithms

In this section, we develop *adaptive iterative strategies* for the *construction* of concrete sequences $(\Lambda_k)_{k\geq 1} \subset \mathcal{F}$ such that the associated Galerkin approximations $\mathbf{u}_{\Lambda_n}$ achieve the optimal, best n-term convergence rates (4.1.8) and (4.1.10). We work in an idealized setting, i.e. under the assumption that Galerkin projection $\mathbf{u}_\Lambda$ and the elliptic problems in D can be computed exactly at a unit cost

The set Λ_k will be generated adaptively. In other words, Λ_k depends on the given datas of the problem, that is the functions ψ_j and f and on the previous solution set Λ_{k-1} through the Galerkin solution $\mathbf{u}_{\Lambda_{k-1}}$. The idea of the algorithms is straightforward, we carry a bulk chasing on the residuals $\mathbf{r}_{\Lambda_k}$ using small values of θ allowing us to have a control on the cardinality. For $\theta \in]0,1[$, we consider the following algorithm

Algorithm 4.4.1

Define $\Lambda_1 := \{0_\mathcal{F}\}$ and compute $\mathbf{u}_{\Lambda_1}$. For $n = 0, 1, \cdots$ do the following:

- *Given that Λ_n has been defined, build $\mathcal{M}_n = \mathcal{M}(\Lambda_n)$ and compute the functions $w_{\Lambda_n,\nu}$ solution of (4.3.4) and theirs norms $c_{\Lambda_n,\nu} := \|w_{\Lambda_n,\nu}\|_V$ for any $\nu \in \mathcal{M}_n$;*
- *Compute the smallest set $\mathcal{S}_n$ in $\mathcal{M}_n$ that satisfies the bulk condition*

$$\sum_{\nu\in\mathcal{S}_n} c^2_{\Lambda_n,\nu} \geq \theta^2 \Big(\sum_{\nu\in\mathcal{M}_n} c^2_{\Lambda_n,\nu} \Big); \tag{4.4.1}$$

- *Enrich Λ_n by adding the element of $\mathcal{S}_n$, i.e. $\Lambda_{n+1} := \Lambda_n \cup \mathcal{S}_n$;*
- *Go to step $n+1$;*

We have the following theorem

Theorem 4.4.2

If $0 < \theta < \min\Big(\kappa(\mathbf{A}), \kappa(\mathbf{A})^{-\frac{1}{2}}\Big)$ and $(\Lambda_k)_{k\geq 0}$ a sequence of set generated by Algorithm 4.4.1, then

$$\|\mathbf{u} - \mathbf{u}_{\Lambda_k}\|_E \leq C\|(c_\nu)\|_{\ell^p(\mathcal{F})}(\#(\Lambda_k)+1)^{-s^*}, \quad s^* = \frac{1}{p} - \frac{1}{2} \tag{4.4.2}$$

with C is a constant depending only on θ.

Proof: First, let us observe that $\theta < \min\left(\kappa(\mathbf{A}), \kappa(\mathbf{A})^{-\frac{1}{2}}\right)$ implies necessarily that $\theta < 1$. The value θ belongs to the range of values for which both Lemma 4.3.3 and Lemma 4.3.4 hold. Following the notation in Algorithm 4.4.1, we denote by $(\mathcal{M}_n)_{n\geq 0}$ the sequence of the margins of the sets Λ_n and by $(\mathcal{S}_n)_{n\geq 0}$ the sequence of intermediate sets i.e. $\mathcal{S}_n = \Lambda_{n+1} \setminus \Lambda_n$. By Lemma 4.3.4, we have a control on the cardinalities of the sets $(\mathcal{S}_n)_{n\geq 0}$ according to

$$\#(\mathcal{S}_n) \leq C_\theta \|\mathbf{r}_{\Lambda_n}\|_S^{-\frac{1}{s^*}} \|(c_\nu)\|_{\ell^p(\mathcal{F})}^{\frac{1}{s^*}},$$

where C_θ is a constant that depends only on θ and $\|\mathbf{A}\|_S$. Using Lemma (4.2.5), we deduce

$$\#(\mathcal{S}_n) \leq C_0 \|\mathbf{u} - \mathbf{u}_{\Lambda_n}\|_E^{-1/s^*} \quad \text{where} \quad C_0 := C_\theta \|\mathbf{A}^{-1}\|^{\frac{1}{2s^*}} \|(c_\nu)\|_{\ell^p(\mathcal{F})}^{\frac{1}{s^*}}.$$

Now from the reduction identity (4.3.11)

$$\|\mathbf{u} - \mathbf{u}_{\Lambda_{k+1}}\|_E \leq \delta \|\mathbf{u} - \mathbf{u}_{\Lambda_k}\|_E, \quad k \geq 0,$$

with $\delta = \sqrt{1 - \frac{\theta^2}{\kappa(\mathbf{A})}}$. we deduce that for any $n \geq 1$ and k in $\{0, ..., n-1\}$

$$\|\mathbf{u} - \mathbf{u}_{\Lambda_n}\|_E \leq \delta^{n-k} \|\mathbf{u} - \mathbf{u}_{\Lambda_k}\|_E.$$

Since $\Lambda_n = \Lambda_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_{n-1}$, then

$$\#(\Lambda_n) = \#(\Lambda_0) + \sum_{k=0}^{n-1} \#(\mathcal{S}_k) \leq 1 + C_0 \sum_{k=0}^{n-1} \|\mathbf{u} - \mathbf{u}_{\Lambda_k}\|_E^{-1/s^*} \leq 1 + C_0 \|\mathbf{u} - \mathbf{u}_{\Lambda_n}\|_E^{-1/s^*} \sum_{k=0}^{n-1} (\delta^{\frac{1}{s^*}})^{n-k},$$

hence $\#(\Lambda_n) \leq 1 + C_1 \|\mathbf{u} - \mathbf{u}_{\Lambda_n}\|_E^{-1/s^*}$, where $C_1 = \frac{C_0}{1-\delta^{\frac{1}{s}}}$, therefore

$$\|\mathbf{u} - \mathbf{u}_{\Lambda_n}\|_E \leq (C_1)^{s^*} (\#(\Lambda_n) - 1)^{-s^*} \leq (2C_1)^{s^*} (\#(\Lambda_n) + 1)^{-s^*},$$

In view of the value of C_0 above, the value $C := \left(\frac{2C_\theta \|\mathbf{A}^{-1}\|^{\frac{1}{2s^*}}}{1-\delta^{\frac{1}{s^*}}}\right)^{s^*}$, is valid for the constant in the inequality (4.4.2) and it only depends on θ, $\|\mathbf{A}\|_S$ and s^*. ■

The previous theorem provides then a positive answer to the objective we have fixed earlier, that is assuming we work in the semi-discrete setting and that the different quantities involved in Algorithm 4.4.1 can be computed exactly, we are able to construct a sequences of nested index sets such that the corresponding Galerkin approximation are near-optimal in the sense of (4.1.8).

Now, we turn to the existence of such sequence of index sets but with the additional lower structure constraint. One natural way to do so is by modifying Algorithm 4.4.1 and adding the constraint that the intermediate set $\mathcal{S}$ is the smallest monotone set in $\mathcal{M}(\Lambda)$ with the bulk property (4.3.10).

For a fixed $0 < \theta < 1$, we consider the following algorithm:

Algorithm 4.4.3

Define $\Lambda_1 := \{0_{\mathcal{F}}\}$ *and compute* $\mathbf{u}_{\Lambda_1}$. *For* $n = 0, 1, \cdots$ *do the following:*

- *Given that* Λ_n *has been defined, build* $\mathcal{M}_n = \mathcal{M}(\Lambda_n)$ *and compute the functions* $w_{\Lambda_n,\nu}$ *solution of* (4.3.4) *and theirs norms* $c_{\Lambda_n,\nu} := \|w_{\Lambda_n,\nu}\|_V$ *for any* $\nu \in \mathcal{M}_n$;
- *Compute the smallest lower set* $\mathcal{S}_n$ *in* $\mathcal{M}_n$ *that satisfies the bulk condition*

$$\sum_{\nu\in\mathcal{S}_n} c^2_{\Lambda_n,\nu} \geq \theta^2 \Big(\sum_{\nu\in\mathcal{M}_n} c^2_{\Lambda_n,\nu} \Big); \tag{4.4.3}$$

- *Enrich* Λ_n *by adding the element of* $\mathcal{S}_n$, *i.e.* $\Lambda_{n+1} := \Lambda_n \cup \mathcal{S}_n$;
- *Go to step* $n + 1$;

The sequences $(\Lambda_k)_{k\geq 1}$ generated by the previous algorithms are lower by construction. Similarly to Theorem 4.4.2, the sequence generated by 4.4.3 is near optimal in the sense of optimality with lower sets (4.1.10). We have in particular

Theorem 4.4.4

If $0 < \theta < \min(\kappa(\mathbf{A}), \kappa(\mathbf{A})^{-\frac{1}{2}})$ *and* $(\Lambda_k)_{k\geq 0}$ *a sequence of nested lower sets generated by Algorithm 4.4.3, then*

$$\|\mathbf{u} - \mathbf{u}_{\Lambda_k}\|_E \leq C \|(c_\nu)\|_{\ell^p_m(\mathcal{F})} (\#(\Lambda_k) + 1)^{-s^*}, \quad s^* = \frac{1}{p} - \frac{1}{2} \tag{4.4.4}$$

with C *the constant in Theorem 4.4.2.*

The result of the theorem can be deduced using the same arguments of the proof of Theorem 4.4.2 and Lemma 4.3.5. The only difference is that $\|(c_\nu)\|_{\ell^p(\mathcal{F})}$ is replaced by $\|(c_\nu)\|_{\ell^p_m(\mathcal{F})}$. One remarks that the arguments of the proof of Theorem 4.4.2 yields the same constant C in both theorems.

Although the two algorithms 4.4.1 and 4.4.3 provide construction of sets that are near optimal in the sense of best n-term approximations (4.1.8) and (4.1.10), they are impractical for several reasons. A first problem is that we can only solve the boundary value problems (4.3.4) approximately, for example using a finite element discretization. We analyze the additional error induced by this discretization in §4.5 . A second problem is that, in our infinite dimensional setting the margins $\mathcal{M}_n$ have infinite cardinality, and therefore there are infinitely many $w_{\Lambda_n,\nu}$ to be computed at each iteration, which requires in principle solving infinitely many boundary value problems. Although this problem does not occur in the finite dimensional setting $d < \infty$, it is still reflected by the fact that the size of $\mathcal{M}_n$ is potentially much larger than that of Λ_n as d gets large

and therefore solving the boundary value problems for all $\nu \in \mathcal{M}_n$ becomes the main source of computational complexity. Finally, the third problem is the computation of the intermediates sets $\mathcal{S}_n$ as the smallest set satisfying the bulk condition in both Algorithm 4.4.1 and Algorithm 4.4.3, for which we do not have an algorithm in linear time.

We propose a solution to the problem of space discretization in §4.5. As for the second problem, we shall exploit the idea of margin truncation used in the previous chapter. The analysis is similar for both algorithms. We restrict our efforts on the modification of Algorithm 4.4.3 which yields near optimal lower sets, the ideas underlying the modification is the same if one consider Algorithm 4.4.1.

4.5 A realistic bulk chasing algorithm

In order to restrict the margins $\mathcal{M}_n$ to finite subsets, we introduce as in Chapter 3 a procedure SPARSE that has the following properties: if Λ is a finite lower set, $\mathcal{M}$ its infinite margin, and $\mathbf{u}_\Lambda$ the Galerkin projection is known, then for any $\eta > 0$,

$$\mathcal{N} := \mathrm{SPARSE}(\Lambda, \mathbf{u}_\Lambda, \eta),$$

is a computable finite subset of $\mathcal{M}$ which is lower in $\mathcal{M}$ and such that

$$\|\mathbf{P}_{\mathcal{M}\setminus\mathcal{N}}\mathbf{r}_\Lambda\|_S = \Big(\sum_{\mathcal{M}\setminus\mathcal{N}} \|w_{\Lambda,\nu}\|_V^2\Big)^{\frac{1}{2}} \leq \eta. \tag{4.5.1}$$

where the functions $w_{\Lambda,\nu}$ are as in Lemma 4.3.1 with the function $\phi_\nu(\Lambda)$ as in (4.3.7) since Λ is lower. One way to construct $\mathcal{N}$ then is by growing incrementally a lower set in $\mathcal{M}$ which corresponds to large values of $\|w_{\Lambda,\nu}\|_V$. However, we do not have a stoping criterion since the norm $\|\mathbf{r}_\Lambda\|_S^2$ of the residual associated with Λ which is the overall sum of contributions $\|w_{\Lambda,\nu}\|_V^2$ is unknown to us. As in Chapter 3, we can show that the SPARSE procedure can be done by activating incrementally new directions. We have the following theorem,

Theorem 4.5.1

Let Λ be a lower set and $\mathcal{M}$ the margin of Λ. For any $\eta > 0$, there exists $\mathcal{N}$ a computable finite lower set in $\mathcal{M}$ such that $\|\mathbf{P}_{\mathcal{M}\setminus\mathcal{N}}\mathbf{r}_\Lambda\|_S \leq \eta$.

Proof: Let $J > 0$ be an integer large enough such that

$$\Big\|\sum_{j>J} |\psi_j|\Big\|_{L^\infty(D)} \leq \frac{\eta}{B}, \tag{4.5.2}$$

where B is a constant that we precise later. We introduce the set $\mathcal{N}_J$ defined by

$$\mathcal{N}_J := \{\nu + e_j : \nu \in \Lambda \text{ and } j \leq J\} \setminus \Lambda. \tag{4.5.3}$$

It is easy to see that $\mathcal{N}_J$ is contained and lower in $\mathcal{M}$. Let us remark also that its definition coincides with the definition (3.4.6) for the set that we used in Chapter 3 for the same purpose with Taylor series. Let now $w_{\Lambda,\nu}$ be the function defined in Lemma 4.3.1. Since Λ is lower, then according to Remark 4.3.2, these functions are the solutions in V of the PDEs

$$-\Delta w = \text{ div } \phi_\nu \text{ in } D, \quad w = 0 \quad \text{on} \quad \partial D,$$

where for each $\nu \in \mathcal{M}$

$$\phi_\nu = \phi_\nu(\Lambda) := \sum_{\substack{j\geq 1 \\ \nu - e_j \in \Lambda}} \beta_{\nu_j - 1} \psi_j \nabla \mathbf{u}_{\Lambda, \nu - e_j}.$$

From theses PDEs and using Green formula, we infer that

$$\|w_{\Lambda,\nu}\|_V^2 = -\int_D \phi_\nu \nabla w_{\Lambda,\nu} = -\sum_{\substack{j\geq 1 \\ \nu - e_j \in \Lambda}} \int_D \beta_{\nu_j - 1} \psi_j \nabla \mathbf{u}_{\Lambda, \nu - e_j} \nabla w_{\Lambda,\nu}, \quad \nu \in \mathcal{M}.$$

Now we make the observation, made also in Chapter 3, that if $\nu \in \mathcal{M} \backslash \mathcal{N}_J$ and $\nu - e_j \in \Lambda$ then necessarily $j > J$. Indeed, if an index ν satisfies $\nu \in \mathcal{M}$ and $\nu - e_j = \nu' \in \Lambda$ with $j \leq J$, then according to the definition of $\mathcal{N}_J$, one has $\nu = \nu' + e_j \in \mathcal{N}_J$. This observation shows in particular that

$$\|w_{\Lambda,\nu}\|_V^2 = -\sum_{\substack{j> J \\ \nu - e_j \in \Lambda}} \int_D \beta_{\nu_j - 1} \psi_j \nabla \mathbf{u}_{\Lambda, \nu - e_j} \nabla w_{\Lambda,\nu}, \quad \nu \in \mathcal{M} \setminus \mathcal{N}_J. \tag{4.5.4}$$

We introduce the function Ψ_J defined on $D \times U$ by

$$\Psi_J(x,y) := \sum_{j>J} y_j \psi_j(x), \quad y \in U, x \in D, \tag{4.5.5}$$

and introduce the operator $\mathbf{T}_J$ defined on $\ell^2(\mathcal{F}, V) \times \ell^2(\mathcal{F}, V)$ by

$$\langle \mathbf{T}_J \mathbf{v}, \mathbf{w} \rangle := \int_U \int_D \Psi_J(x,y) \nabla v \nabla w dx dy,$$

where v and w are the function in $\mathcal{V}_2$ with representations $\mathbf{v}$ and $\mathbf{w}$ in $\ell^2(\mathcal{F}, V)$. If $\mathbf{v}$ supported in Λ and $\mathbf{w}$ supported in $\mathcal{M}$, then by the same arguments used in the proof of Lemma 4.2.1, we obtain

$$\langle \mathbf{T}_J \mathbf{v}, \mathbf{w} \rangle = \sum_{\nu \in \mathcal{M}} \sum_{\nu' \in \Lambda} \sum_{j>J} \int_D \Big(\int_U y_j L_\nu L_{\nu'} dy \Big) \psi_j \nabla \mathbf{v}_{\nu'} \nabla \mathbf{w}_\nu dx = \sum_{\nu \in \mathcal{M}} \sum_{j>J} \int_D \beta_{\nu - e_j} \psi_j \nabla \mathbf{v}_{\nu - e_j} \nabla \mathbf{w}_\nu dx,$$

We introduce the notation $\mathbf{w}^{\mathcal{N}_J}$ for the vector supported in $\mathcal{M} \setminus \mathcal{N}_J$ and have coordinates $w_{\Lambda,\nu}$ for each $\nu \in \mathcal{M} \setminus \mathcal{N}_J$. By (4.3.8), we have $\|\mathbf{P}_{\mathcal{M} \setminus \mathcal{N}_J} \mathbf{r}_\Lambda\|_S^2 = \|\mathbf{w}^{\mathcal{N}_J}\|^2$.

Moreover, summing the formulas (4.5.4) for the indices $\nu \in \mathcal{M} \setminus \mathcal{N}_J$ and using the above equality with $\mathbf{v} = \mathbf{u}_\Lambda$ and $\mathbf{w} = \mathbf{w}^{\mathcal{N}_J}$, we infer

$$\|\mathbf{w}^{\mathcal{N}_J}\|^2 = \langle \mathbf{T}_J \mathbf{u}_\Lambda, \mathbf{w}^{\mathcal{N}_J} \rangle, \tag{4.5.6}$$

which implies

$$\|\mathbf{P}_{\mathcal{M}\setminus\mathcal{N}} \mathbf{r}_\Lambda\|_S = \|\mathbf{w}^{\mathcal{N}_J}\| \leq \|\mathbf{T}_J\|_S \|\mathbf{u}_\Lambda\|. \tag{4.5.7}$$

It is easy to see that the optimality of the Galerkin projection $\mathbf{u}_\Lambda$ implies $\|\mathbf{u}_\Lambda\|_E^2 \leq \|\mathbf{u}\|_E^2$, therefore $\|\mathbf{u}_\Lambda\| \leq 2\sqrt{R/r}\|\mathbf{u}\|$. Also, elementary arguments show that

$$\|\mathbf{T}_J\|_S \leq \|\Psi_J\|_{L^\infty(D\times U)} \leq \left\| \sum_{j>J} |\psi_j| \right\|_{L^\infty(D)}.$$

We deduce that

$$\|\mathbf{P}_{\mathcal{M}\setminus\mathcal{N}} \mathbf{r}_\Lambda\|_S \leq 2\sqrt{R/r} \frac{\|f\|_{V^*}}{r} \left\| \sum_{j>J} |\psi_j| \right\|_{L^\infty(D)} \leq \eta,$$

if B is the constant $\frac{1}{2\sqrt{R/r}} \frac{r}{\|f\|_{V^*}}$, which completes the proof. ■

Now, we are able to define the practical algorithms with similar techniques used in Chapter 3. For the sake of clarity, given a lower set Λ and $\mathcal{S}$ a subset of its margin $\mathcal{M}$, we introduce the notation

$$e_\Lambda(\mathcal{S}) := \|\mathbf{P}_\mathcal{S} \mathbf{r}_\Lambda\|_S^2 = \sum_{\nu\in\mathcal{S}} \|w_{\Lambda,\nu}\|_V^2, \quad \mathcal{S} \subset \mathcal{M}(\Lambda). \tag{4.5.8}$$

The quantiy $e_\Lambda(\mathcal{S})$ is the contribution of $\mathcal{S}$ to the norm of the residual. It is to be compared with

$$e(\mathcal{S}) := \sum_{\nu\in\mathcal{S}} \|t_\nu\|_a^2 \tag{4.5.9}$$

the energy associated with $\mathcal{S}$ of Taylor coefficients as defined in Chapter 3. The previous lemma does not implies that $\mathcal{N}_J$ captures directly a fraction of the energy $e_\Lambda(\mathcal{M}) = \|\mathbf{r}_\Lambda\|_S$. We propose to use an incremental strategy

$$(\mathcal{N}, \eta) := \text{OVERGROW}(\mathcal{M}, \mathbf{u}_\Lambda, \theta), \tag{4.5.10}$$

as in Chapter 3, which giving $\mathcal{M}$ the margin of Λ, output the value η and a finite restricted margin $\mathcal{N}$ such that $e_\Lambda(\mathcal{M}\setminus\mathcal{N}) \leq \eta$ and captures at least a fraction θ of the energy $e_\Lambda(\mathcal{M}) = \|\mathbf{r}_\Lambda\|_S^2$. For example, using the restricted margin, this can be done by incrementing J, and accordingly growing $\mathcal{N}_J$ until we captures the desired fraction.

Algorithm 4.5.2

Let Λ a lower set, $\mathcal{M}$ its margin, $\theta \in]0,1[$ and $\eta > 0$. Let $j = 0$, then do the following

- *Define* $\eta_j := 2^{-j}\eta$ *and* $\mathcal{M}_j := \mathrm{SPARSE}(\Lambda, \mathbf{u}_\Lambda, \eta_j)$;
- *Compute* $w_{\Lambda,\nu}$ *and* $c_{\Lambda,\nu}$ *for* $\nu \in \mathcal{M}_j$ *and then the quadratic sum* $e_\Lambda(\mathcal{M}_j)$;
- *If* $e_\Lambda(\mathcal{M}_j) < \frac{2(2-\theta^2)}{1-\theta^2}\eta_j^2$, *then go directly to step* $j+1$;
- *Else, terminate the loop in* j, *and output the set* $\mathcal{M}_j$ *and the value* η_j.

We have

$$e_\Lambda(\mathcal{M}_j) \geq \theta^2 e_\Lambda(\mathcal{M}) = \theta^2 \|\mathbf{r}_\Lambda\|_S^2 \tag{4.5.11}$$

Proof: The previous loop always terminates, indeed, η_j decrease to 0, while the energies of the residual on the restricted margin $e(\mathcal{M}_j)$ increases. Let J be the last integer in the previous loop. On the one hand $e_\Lambda(\mathcal{M}_J) \geq \frac{2(2-\theta^2)}{1-\theta^2}\eta_J^2$, therefore

$$e(\mathcal{M}_J) \geq \theta^2 e(\mathcal{M}_J) + 2(2-\theta^2)\eta_J^2 \geq \theta^2 e(\mathcal{M}_J) + (2-\theta^2)\eta_J^2.$$

One the other hand from the defition of $\mathcal{M}_J$, we have $e_\Lambda(\mathcal{M} \setminus \mathcal{M}_J) \leq \eta_J^2$, it follows that $e_\Lambda(\mathcal{M}_J) \geq e_\Lambda(\mathcal{M}) - \eta_J^2$, hence

$$e_\Lambda(\mathcal{M}_J) \geq \theta^2(e_\Lambda(\mathcal{M}) - \eta_J^2) + (2-\theta^2)\eta_J^2 \geq \theta^2 e_\Lambda(\mathcal{M}) + 2(1-\theta^2)\eta_J^2 \geq \theta^2 e_\Lambda(\mathcal{M}),$$

which finishes the proof. ■

We are able to capture a bulk of the residual by a finite lower set in $\mathcal{M}$. This allows us to be delivered form the serious constraint of infinite cardinality of the set controlling the energy. As in Chapter 3, we can now propose a realistic bulk chasing algorithm. Fixing $0 < \theta < 1$, and consider the following algorithm.

Algorithm 4.5.3

Define $\Lambda_0 := \{0_\mathcal{F}\}$, *compute* $\mathbf{u}_{\Lambda_0}$, *and set* $\eta_0 = c_{0_\mathcal{F}}$. *For the values* $n = 0, 1, \cdots$, *do the following*

- *Given that* Λ_n *has been defined and* $\mathbf{u}_{\Lambda_n}$ *has be computed define* $\mathcal{M}_n = \mathcal{M}(\Lambda_n)$.
- *Output the restricted margin* $(\mathcal{M}_{j_n}, \eta_{j_n}) := \mathrm{OVERGROW}(\mathcal{M}_n, \mathbf{u}_{\Lambda_n}, \theta)$, *and define* $\eta_{n+1} := \eta_{j_n}$.
- *Enrich* Λ_n *by* $\mathcal{S}_n$ *the smallest lower set in* $\mathcal{M}_{j_n}$ *such that*

$$e_\Lambda(\mathcal{S}_n) \geq \frac{1}{4}\, e_\Lambda(\mathcal{M}_{j_n});$$

- *Go to step* $n+1$;

At every step of the algorithm, we have $e_\Lambda(\mathcal{S}_n) \geq \frac{1}{4} e_\Lambda(\mathcal{M}_{j_n}) \geq \frac{\theta^2}{4} e(\mathcal{M}_n)$, therefore $\|\mathbf{P}_{\mathcal{S}_n} \mathbf{r}_{\Lambda_n}\|_S \geq \frac{\theta}{2} \|\mathbf{r}_{\Lambda_n}\|_S$. In view of (4.3.11)

$$\|\mathbf{u} - \mathbf{u}_{\Lambda_{n+1}}\|_E \leq \delta_2 \|\mathbf{u} - \mathbf{u}_\Lambda\|_E, \tag{4.5.12}$$

where $\delta = \sqrt{1 - \frac{\theta^2}{4\kappa(\mathbf{A})}}$. Using exactly the same arguments of the results following the impractical algorithm 4.4.3, with the only difference of $\theta/2$ replacing θ, we can prove the following theorem

Theorem 4.5.4

If $0 < \theta/2 < \min(\kappa(\mathbf{A}), \kappa(\mathbf{A})^{-\frac{1}{2}})$ and $(\Lambda_n)_{n\geq 0}$ a sequence of nested monotone set generated by 4.5.3, then

$$\|\mathbf{u} - \mathbf{u}_{\Lambda_n}\|_E \leq C \|(c_\nu)\|_{\ell^p_m(\mathcal{F})} (\#(\Lambda_n) + 1)^{-s^*}, \quad s^* = \frac{1}{p} - \frac{1}{2} \tag{4.5.13}$$

with C the constant in Theorem 4.4.2.

As discussed is Chapter 3, the procedure SPARSE and OVERGROW might produce large restricted margins. 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 $\|w_{\Lambda,\nu}\|_V$.

4.6 Space discretization

The previous convergence results are benchmarks as to how well the functions $u(y)$ may be jointly approximated in the mean square sense with a prescribed accuracy by a finite linear combination

$$\sum_{\nu \in \Lambda_k^{L*}} v_\nu L_\nu, \quad \text{or} \quad u_\Lambda := \sum_{\nu \in \Lambda_k} u_{\Lambda,\nu} L_\nu.$$

These results are *semidiscrete* in that any numerical realization of such a linear combination would itself involve the approximation of the Legendre coefficient v_ν or the coordinates of Galerkin projection $u_{\Lambda,\nu} \in V$ through discretization in D, such as for example by the Finite Element method in D.

Specifically, we consider the approximation of the functions in $V = H_0^1(D)$ with D a bounded Lipschitz polyhedron D by a one parameter affine family of continuous piecewise linear Finite Element spaces $(V_h)_{h>0}$ on a shape regular family of simplicial triangulations of mesh-width $h > 0$ in the sense of [28] (higher order, iso-parametric Finite Element families in curved domains could equally be considered; we confine our analysis to affine, piecewise linear Finite Element families for ease of exposition

only). Convergence rates of such Finite Element approximations are determined by the regularity of the functions being approximated in D. For this, further regularity assumptions on f are required. Again for ease of exposition, we shall assume $f \in L^2(D) \subset V^*$. Then

$$\|f\|_{V^*} \leq C_P \|f\|_{L^2(D)}, \tag{4.6.1}$$

where C_P is the Poincaré constant of D (i.e. $C_P = 1/\sqrt{\lambda_1}$ with λ_1 being the smallest eigenvalue of the Dirichlet Laplacian in D). Then the smoothness space $W \subset V$ is the space of all solutions to the Dirichlet problem

$$-\Delta u = f \quad \text{in} \quad D, \qquad u = 0 \quad \text{on} \quad \partial D, \tag{4.6.2}$$

with $f \in L^2(D)$, that is

$$W = \Big\{ v \in V : \Delta v \in L^2(D) \Big\}. \tag{4.6.3}$$

We define the W-(semi) norm and the W-norm by

$$|v|_W = \|\Delta v\|_{L^2(D)}, \qquad \|v\|_W := \|v\|_V + |v|_W. \tag{4.6.4}$$

It is well-known that $W = H^2(D) \cap V$ for convex $D \subset \mathbb{R}^m$. Then any $w \in W$ may be approximated in V with convergence rate $\mathcal{O}(h)$ by continuous, piecewise linear Finite Element approximations on regular quasi-uniform simplicial partitions of D of meshwidth h (cf. e.g. [28, 15]). Therefore, denoting $M = \dim(V_h) \sim h^{-m}$ the dimension of the Finite Element space, we have for all $w \in W$ the convergence rate

$$\inf_{v_h \in V_h} \|w - v_h\|_V \leq C M^{-\frac{1}{m}} |w|_W. \tag{4.6.5}$$

More generally, for non-convex polyhedra, the space W is not contained in $H^2(D)$, and the convergence rate as $M = \dim(V_h) \to \infty$ is reduced to

$$\inf_{v_h \in V_h} \|w - v_h\|_V \leq C_t M^{-t} |w|_W. \tag{4.6.6}$$

with some $0 < t < \frac{1}{m}$.

The discretised solution map u_h belongs to $\mathcal{V}_{h,2} = L^2(U, V_h, d\varrho)$ and it is the unique solution of the variational problem

$$\mathcal{B}(u_h, v_h) = \mathcal{L}(v_h), \qquad v_h \in \mathcal{V}_{h,2}, \tag{4.6.7}$$

where we have defined the bilinear form $\mathcal{B}$ over $\mathcal{V}_{h,2} \times \mathcal{V}_{h,2}$ and the linear form $\mathcal{L}$ over $\mathcal{V}_{h,2}$ as in (4.2.2). In particular, the integrals over D are understood as extensions by continuity from $L^2(D) \times L^2(D)$ to duality pairings between V_h^* and V_h. Using exactly the same analysis of the previous sections, one reformulates the problem as a system

$$\mathbf{A}\mathbf{u}_h = \mathbf{f} \tag{4.6.8}$$

where $\mathbf{A}$ is an operator from $\ell^2(\mathcal{F}, V_h)$ into $\ell^2(\mathcal{F}, V_h^*)$ with entries as in Lemma 4.2.1 defined over $V_h \times V_h$ and $\mathbf{u}_h$ and $\mathbf{f}$ the sequence of Legendre coefficients of u_h and f in the basis $(L_\nu)_{\nu\in\mathcal{F}}$. Assuming that the coordinates of Galerkin projections $\mathbf{u}_{h,\Lambda}$ can be computed exactly in V_h and that also the associated functions $w_{h,\Lambda,\nu}$ can be computed exactly in V_h, the various bulk chasing algorithms yields approximation $\mathbf{u}_{h,\Lambda}$ to $\mathbf{u}_h$ with the rates

$$\|\mathbf{u}_h - \mathbf{u}_{h,\Lambda_n}\|_E \leq C\|(v_{h,\nu})\|_{\ell^p_m(\mathcal{F})}(\#(\Lambda_n)+1)^{-s}, \quad s = \frac{1}{p} - \frac{1}{2}. \tag{4.6.9}$$

One only needs to quantify the additional discretization error $\|\mathbf{u}-\mathbf{u}_h\|_E$. We have that

$$\|\mathbf{u} - \mathbf{u}_h\|_E \leq \sqrt{R}\sup_{y\in U}\|u(y) - u_h(y)\|_V \tag{4.6.10}$$

4.7 Approximation of Galerkin Projection

In the previous sections, we have assumed that, given an index set Λ, we are able to compute the Galerkin approximation $\mathbf{u}_\Lambda$ of $\mathbf{u}$ or $\mathbf{u}_{h,\Lambda}$ of $\mathbf{u}_h$ exactly and in unit cost in $\ell^2(\Lambda, V)$ and $\ell^2(\Lambda, V_h)$ respectively. This is obviously not possible in practice. The Galerkin projection can only be computed to a desired accuracy and the cost of the approximation depends on this target accuracy and possibly on the size of the index set Λ.

In this section, we show that the exact Galerkin projection $\mathbf{u}_\Lambda$ and the discrete Galerkin projection $\mathbf{u}_{\Lambda,h}$ can be approximated to any given accuracy ε using iterative Jacobi method. We then investigate how the adaptive algorithm can be modified in order to take into account this approximation. Since the semi-discrete and fully-discrete settings are similar in the sense they can be treated similarly if a unique space V_h is used for discretization of all the function in $V = H_0^1(D)$, we only consider the semi-discrete setting and we work only with the space V.

4.7.1 Iterative Jacobi Method

The key point in the analysis of this section is that the uniform ellipticity assumption implies that the matrix of operators $\mathbf{A}$ is *"Diagonally dominant"*. For the sake of notational simplicity, we consider in this section the space $\ell^2(\Lambda, V)$ defined in (4.2.21) to be the space of V-valued sequences indexed in Λ which are square integrable, i.e.

$$\ell^2(\Lambda, V) := \left\{\mathbf{v} = (\mathbf{v}_\nu)_{\nu\in\Lambda} : \sum_{\nu\in\Lambda}\|\mathbf{v}_\nu\|_V^2 < \infty\right\}. \tag{4.7.1}$$

Remark that if Λ is finite, $\ell^2(\Lambda, V)$ is merely the space of V-valued sequences indexed in Λ. Given $\Lambda \subset \mathcal{F}$ a set of indices, we introduce the notations $\mathbf{A}_\Lambda := (\mathbf{A}_{\nu\nu'})_{\nu,\nu'\in\Lambda}$

and $\mathbf{f}_\Lambda := (\mathbf{f}_\nu)_{\nu\in\Lambda}$ for the sections of the infinite matrix of operators $\mathbf{A}$ and the infinite vector $\mathbf{f}$ restricted to the index set Λ. The matrix of operators $\mathbf{A}_\Lambda$ can be seen as an operator from $\ell^2(\Lambda;V)$ into $\ell^2(\Lambda;V^*)$. Indeed, for $\mathbf{v},\mathbf{w}\in\ell^2(\Lambda;V)$, we have

$$\langle \mathbf{A}_\Lambda \mathbf{v},\mathbf{w}\rangle = \langle \mathbf{A}\mathbf{v},\mathbf{w}\rangle, \tag{4.7.2}$$

where with a slight abuse $\mathbf{v}$ and $\mathbf{w}$ in the right side are in $\ell^2(\mathcal{F},V)$ with null elements for indices $\nu\notin\Lambda$. The duality products $\langle\cdot,\cdot\rangle$ are considered with between $\ell^2(\Lambda;V*)$ and $\ell^2(\Lambda;V^*)$ and between $\ell^2(\mathcal{F};V^*)$ and $\ell^2(\mathcal{F};V)$ respectively, as in (4.2.15).

Since the Galerkin approximation $\mathbf{u}_\Lambda$ is supported in Λ, we may consider it as a vector in the newly defined space $\ell^2(\Lambda;V)$. We have then that $\mathbf{u}_\Lambda$ is the unique solution to

$$\mathbf{A}_\Lambda \mathbf{u}_\Lambda = \mathbf{f}_\Lambda. \tag{4.7.3}$$

Since $\mathbf{f}$ is support in $\{0_\mathcal{F}\}$, then unless the null multi-index $0_\mathcal{F}$ belongs to Λ, the previous system is trivial.

We suppose in the sequel that $0_\mathcal{F}\in\Lambda$. As we suggested earlier, the key point for the convergence of the Jacobi iterative method is that the section $\mathbf{A}_\Lambda$ is diagonally dominant. To see this, we introduce diagonal matrix $\bar{\mathbf{A}}_\Lambda$ of $\mathbf{A}_\Lambda$ defined in the obvious way by, $\bar{\mathbf{A}}_\Lambda := (\bar{\mathbf{A}}_{\nu\nu'})_{\nu,\nu'\in\Lambda}$, with

$$(\bar{\mathbf{A}}_\Lambda)_{\nu\nu'} = \delta_{\nu,\nu'}\mathbf{A}_{\nu\nu'}, \quad \nu,\nu'\in\Lambda, \tag{4.7.4}$$

and the matrix $\Psi_\Lambda := \mathbf{A}_\Lambda - \bar{\mathbf{A}}_\Lambda$. The explicit formulas of the operators $\mathbf{A}_{\nu\nu}$, given in Lemma 4.2.1, implies: for any $\mathbf{v},\mathbf{w}\in\ell^2(\Lambda;V)$ and v and w the corresponding function in $\mathcal{V}_2$

$$\langle \bar{\mathbf{A}}_\Lambda\mathbf{v},\mathbf{w}\rangle = \int_U\int_D \bar{a}(x)\nabla v\nabla w, \quad \text{hence} \quad \langle \Psi_\Lambda\mathbf{v},\mathbf{w}\rangle = \int_U\int_D \Big(a(x,y)-\bar{a}(x)\Big)\nabla v\nabla w. \tag{4.7.5}$$

The uniform ellipticity assumption $\mathbf{UEA}(r,R)$ applied at $y=0$ implies

$$0<r\le \bar{a}(x)\le R<\infty, \qquad x\in D, \tag{4.7.6}$$

therefore the operator $\bar{\mathbf{A}}_\Lambda$ define a boundedly invertible operator from $\ell^2(\Lambda,V)$ into $\ell^2(\Lambda,V^*)$. The operator $\bar{\mathbf{A}}_\Lambda$ induces a norm on $\ell^2(\Lambda,V)$ defined by

$$\|\mathbf{v}\|_{\bar{a}} := \sqrt{\langle \bar{\mathbf{A}}_\Lambda\mathbf{v},\mathbf{v}\rangle}, \qquad \mathbf{v}\in\ell^2(\Lambda,V), \tag{4.7.7}$$

This norm is equivalent to $\|\cdot\|$ and $\|\cdot\|_E$. We have for any $y\in U$ that

$$|a(x,y)-\bar{a}(x)|\le\sum_{j\ge1}|\psi_j(x)|\le\bar{a}(x)-r\le\gamma\bar{a}(x), \qquad x\in D, \tag{4.7.8}$$

where $\gamma = 1-\frac{r}{R}<1$. Therefore, given $\mathbf{v},\mathbf{w}\in\ell^2(\Lambda,V)$ and v and w as above, we obtain by Cauchy-Schwartz formula

$$|\langle\Psi_\Lambda\mathbf{v},\mathbf{w}\rangle|\le\gamma\|\mathbf{v}\|_{\bar{a}}\|\mathbf{w}\|_{\bar{a}}. \tag{4.7.9}$$

Now, we are able to give the following result

Lemma 4.7.1

Let $(\mathbf{u}^n)_{n\geq 0}$ be a sequence in $\ell^2(\Lambda, V)$ defined by

$$\mathbf{u}^0 \in \ell^2(\Lambda; V), \quad \bar{\mathbf{A}}_\Lambda \mathbf{u}^{n+1} = \mathbf{f}_\Lambda - \Psi_\Lambda \mathbf{u}^n, \quad n \geq 0.$$

This sequence satisfies

$$\|\mathbf{u}^n - \mathbf{u}_\Lambda\|_{\bar{a}} \leq \gamma^n \|\mathbf{u}^0 - \mathbf{u}_\Lambda\|_{\bar{a}},$$

with $\gamma = 1 - \frac{r}{R}$.

Proof: Since $\mathbf{u}_\Lambda$ is the unique solution to $\mathbf{A}\mathbf{u}_\Lambda = \mathbf{f}_\Lambda$, then $\bar{\mathbf{A}}_\Lambda \mathbf{u}_\Lambda = (\mathbf{f}_\Lambda - \Psi_\Lambda \mathbf{u}_\Lambda)$. Therefore for any $n \geq 0$

$$\bar{\mathbf{A}}_\Lambda(\mathbf{u}^{n+1} - \mathbf{u}_\Lambda) = -\Psi_\Lambda(\mathbf{u}^n - \mathbf{u}_\Lambda), \tag{4.7.10}$$

which combined with (4.7.9) implies for any $n \geq 0$

$$\|\mathbf{u}^{n+1} - \mathbf{u}_\Lambda\|_{\bar{a}}^2 \leq \gamma \|\mathbf{u}^{n+1} - \mathbf{u}_\Lambda\|_{\bar{a}} \|\mathbf{u}^n - \mathbf{u}_\Lambda\|_{\bar{a}}. \tag{4.7.11}$$

The proof can then be completed by an immediate induction on $n \geq 0$. ■

Remark 4.7.2

The principal of the iterative Jacobi method consists in writing the matrix $\mathbf{A}$ *as a sum* $\mathbf{A} = \mathbf{D} + \mathbf{M}$ *with* $\mathbf{D}$ *invertible and* $\rho(\mathbf{D}^{-1}\mathbf{M}) < 1$. *The previous analysis suggest that one can always find such decomposition in our present setting by only constructing* $\mathbf{D}$ *by a similar construction of* $\mathbf{A}$ *but with* $\bar{a}(x)$ *is used instead of* $a(x, y)$ *and setting* $\mathbf{M} = \mathbf{A} - \mathbf{D}$. *We should also point out that this construction is independent of the choice of the Legendre polynomials as the polynomials basis for* $\mathbb{P}_\Lambda$. *The choice of the Legendre polynomials is rather motivated by the sparsity of the matrix* $\mathbf{A}$ *obtained and analytic regularity that allows us to find a near optimal sets* Λ *using these polynomials, see Chapter 1.*

Remark 4.7.3

Given two index sets $\Lambda \subset \tilde{\Lambda}$ *and assuming we know the Galerkin projection* $\mathbf{u}_\Lambda$, *one may choose for the iterative computation of* $\mathbf{u}_{\tilde{\Lambda}}$ *the initial guess* $\mathbf{u}^0$ *to be the vector in* $\ell^2(\tilde{\Lambda}, V)$ *that coincides with* $\mathbf{u}_\Lambda$ *for the indices* $\nu \in \Lambda$ *and has null coordinates for the indices* $\nu \in \tilde{\Lambda} \setminus \Lambda$. *For this choice, we have*

$$\|\mathbf{u}^0 - \mathbf{u}_{\tilde{\Lambda}}\|_{\bar{a}} = \|\mathbf{u}_\Lambda - \mathbf{u}_{\tilde{\Lambda}}\|_{\bar{a}} \lesssim \|\mathbf{u}_\Lambda - \mathbf{u}_{\tilde{\Lambda}}\|_E \leq \|\mathbf{u} - \mathbf{u}_\Lambda\|_E \tag{4.7.12}$$

where we have used the optimality of the Galerkin projection $\mathbf{u}_{\tilde{\Lambda}}$;

$$\|\mathbf{u} - \mathbf{u}_\Lambda\|_E^2 = \|\mathbf{u} - \mathbf{u}_{\tilde{\Lambda}}\|_E^2 + \|\mathbf{u}_{\tilde{\Lambda}} - \mathbf{u}_\Lambda\|_E^2. \tag{4.7.13}$$

Therefore, the sequence $(\mathbf{u}^n)_{n\geq 1}$ *computed by the iterative algorithm for the approximation of* $\mathbf{u}_{\tilde{\Lambda}}$ *satisfies*

$$\|\mathbf{u}^n - \mathbf{u}_{\tilde{\Lambda}}\|_{\bar{a}} \lesssim \gamma^n \|\mathbf{u} - \mathbf{u}_\Lambda\|_E, \quad n \geq 0. \tag{4.7.14}$$

In particular, if $\Lambda = \Lambda_k$ and $\tilde{\Lambda} = \Lambda_{k+1}$ are two set output by the adaptive algorithms described earlier, then

$$\|\mathbf{u}^n - \mathbf{u}_{\Lambda_{k+1}}\|_{\bar{a}} \lesssim \gamma^n C \|(c_\nu)\|_{\ell^p_{m(\mathcal{F})}} (\#(\Lambda_k)+1)^{-s^*}, \quad n \geq 0. \tag{4.7.15}$$

One can then decide to stop the iterative computation of the Galerkin approximation $\mathbf{u}_{\Lambda_{k+1}}$ after a number of iterations $N = N(k)$ such that

$$\gamma^N (\#(\Lambda_k)+1)^{-s^*} \leq (\#(\Lambda_{k+1})+1)^{-s^*}. \tag{4.7.16}$$

Let us return to the iterative algorithm. We have a non empty set Λ containing $0_{\mathcal{F}}$ and we want to compute the sequence $(\mathbf{u}^n)_{n\geq 0}$ that eventually converges to $\mathbf{u}_\Lambda$. We shall explain more explicitly how the vector $\mathbf{u}^{n+1}$ can be deduced from the vector $\mathbf{u}^n$. We introduce the notation $\mathbf{u}^n_\nu$, $\nu \in \Lambda$ for the coordinates of the vector $\mathbf{u}^n$. Since for any $\nu, \nu' \in \mathcal{F}$, we have $(\mathbf{A}_\Lambda)_{\nu,\nu'} = \mathbf{A}_{\nu,\nu'} = 0$ unless $\nu = \nu'$ or $\nu = \nu' \pm e_j$ for some $j \geq 1$, then the formula $\bar{\mathbf{A}}_\Lambda \mathbf{u}^{n+1} = \mathbf{f}_\Lambda - \Psi_\Lambda \mathbf{u}^n$ is equivalent to

$$\mathbf{A}_{\nu\nu}\mathbf{u}^{n+1}_\nu = \mathbf{f}_\nu - \sum_{j\geq 1:\nu-e_j\in\Lambda} \mathbf{A}_{\nu,\nu-e_j}\mathbf{u}^n_\nu - \sum_{j\geq 1:\nu+e_j\in\Lambda} \mathbf{A}_{\nu,\nu+e_j}\mathbf{u}^n_\nu, \quad \nu\in\Lambda. \tag{4.7.17}$$

Using the explicit formulas of the operators $\mathbf{A}_{\nu,\nu'}$ given in Lemma 4.2.1, we deduce that the coordinates u^{n+1}_ν, $\nu \in \Lambda$ are the unique solutions in V of the following variational formulas:

$$\int_D \bar{a}\nabla \mathbf{u}^{n+1}_0 \nabla w = \int_D fw - \sum_{j\geq 1:e_j\in\Lambda} \beta_0 \int_D \psi_j \nabla \mathbf{u}^n_{e_j} \nabla w, \quad w \in V, \tag{4.7.18}$$

and for any $\nu \in \Lambda - \{0\}$

$$\int_D \bar{a}\nabla \mathbf{u}^{n+1}_\nu \nabla w = -\sum_{j\geq 1:\nu+e_j\in\Lambda} \beta_{\nu_j} \int_D \psi_j \nabla \mathbf{u}^n_{\nu+e_j}\nabla w - \sum_{j\geq 1:\nu-e_j\in\Lambda} \beta_{\nu_j-1} \int_D \psi_j \nabla \mathbf{u}^n_{\nu-e_j}\nabla w, \quad w\in V. \tag{4.7.19}$$

The sequence $(\beta_n)_{n\geq 0}$ is defined in Lemma 4.2.1. As with the recursion formula (3.2.3) used in Chapter 3 for the computation of Taylor coefficients, we remark that determining u^{n+1}_Λ using u^n_Λ requires the successive numerical solution of the same "nominal" elliptic problems (4.7.18) with $\#(\Lambda)$ many right hand sides. In particular, in order to compute a numerical approximation of the $(u^{n+1}_{\Lambda,\nu})_{\nu\in\Lambda}$, *a single discretized, parameter-independent "nominal" elliptic problem (4.7.18) in the domain D must be solved* with $\#(\Lambda)$ many load cases.

Unlike the Taylor coefficients, the recursive formulas for the computation of u^{n+1}_ν for $\nu \in \Lambda$ depends also on the indices $\nu + e_j$. We should note that the numbers of such indices does not exceed $\#(\mathrm{supp}(\Lambda))$, which is smaller than the parametric dimension d and than $\#(\Lambda)$.

We should finally remark that when $\Lambda = \{0_{\mathcal{F}}\}$, the Galerkin solution $\mathbf{u}_\Lambda$ is merely the vector with one coordinates $\mathbf{u}_{\Lambda,0_{\mathcal{F}}}$ associated with the Legendre polynomial $L_{0_{\mathcal{F}}} = 1$ satisfying $\mathbf{A}_{0_{\mathcal{F}},0_{\mathcal{F}}}\mathbf{u}_{\Lambda,0_{\mathcal{F}}} = f$. This implies that

$$u_{\{0_{\mathcal{F}}\}} := u(0). \tag{4.7.20}$$

A similar straightforward computation holds also if Λ is a rectangular block as we have already mentioned in the general introcution, formula (3.24).

4.7.2 An adaptive algorithms with approximate Galerkin projection

In the previous sections, we have proposed adaptive algorithms where we suppose the knowledge of $\mathbf{u}_\Lambda$ for any set Λ in $\mathcal{F}$. In fact, we can only approximate $\mathbf{u}_\Lambda$ to any given accuracy using for instance iterative methods as explained in §4.7. We suppose then that given Λ lower, we can compute $\mathbf{u}^\varepsilon_\Lambda$ an approximate to $\mathbf{u}_\Lambda$ to any given accuracy ε, i.e.

$$\|\mathbf{u}^\varepsilon_\Lambda - \mathbf{u}_\Lambda\| \leq \varepsilon \tag{4.7.21}$$

Our objective is to design adaptive algorithm of the type proposed in sections 3.4 and 3.5 where at each iteration n we compute an approximate $\mathbf{u}^{\varepsilon_n}_{\Lambda_n}$ of the Galerkin approximation $\mathbf{u}_{\Lambda_n}$ yet the approximation of $\mathbf{u}$ by theses approximate Galerkin projections is near-optimal in the sense of (4.1.8) and (4.1.10) up to a controlled numerical error.

Rather than striving for utmost generality, we only focus on adjusting the algorithm 4.4.3 that constructs near optimal lower sets in the sense of (4.1.10) and assume we work on the finite dimension setting $d < \infty$, so that we ignore the problem of margin truncation. At every step of the algorithm, we will only be able to compute $\mathbf{u}^{\varepsilon_n}_{\Lambda_n}$ an approximate of $\mathbf{u}_{\Lambda_n}$ to any target accuracy ε_n. We propose to keep to some extent the same simple features of the algorithm.

Given a lower set Λ for which an approximate $\mathbf{u}^\varepsilon_\Lambda$ in the sense of (4.7.21) is known and $\mathcal{M}$ its margin, we introduce the notations $w^\varepsilon_{\Lambda,\nu}$ for the solutions in V of the systems

$$-\Delta w = \text{ div } \phi^\varepsilon_\nu \quad \text{in} \quad D, \quad w_{|\partial D} = 0, \tag{4.7.22}$$

with for each $\nu \in \mathcal{M}$

$$\phi^\varepsilon_\nu = \phi^\varepsilon_\nu(\Lambda) := \sum_{\substack{j\geq 1 \\ \nu - e_j \in \Lambda}} \beta_{\nu_j - 1}\psi_j \nabla \mathbf{u}^\varepsilon_{\Lambda,\nu-e_j}. \tag{4.7.23}$$

where $\mathbf{u}^\varepsilon_{\Lambda,\nu}$, are the coordinates of $\mathbf{u}^\varepsilon_\Lambda$ and the sequence $(\beta_n)_{n\geq 0}$ is as in Lemma 4.2.1. The functions $w^\varepsilon_{\Lambda,\nu}$ imitate the functions $w_{\Lambda,\nu}$ defined similarly in Lemma 4.3.1 and Remark 4.3.7 and used to quantify the contribution to the energy of residual $\mathbf{r}_\Lambda$ supported

in every $\nu \in \mathcal{M}$. However it should be noted that they do not play the same role for the "approximate residual" $\mathbf{r}^\varepsilon_\Lambda := \mathbf{f} - \mathbf{A}\mathbf{u}^\varepsilon_\Lambda$. Now we introduce the notation

$$e^\varepsilon_\Lambda(\mathcal{S}) := \sum_{\nu\in\mathcal{S}} \|w^\varepsilon_{\Lambda,\nu}\|^2_V, \quad \mathcal{S} \subset \mathcal{M}. \tag{4.7.24}$$

This notation is the counterpart of the notation $e_\Lambda(\mathcal{S}) = \|\mathbf{P}_\mathcal{S}\mathbf{r}_\Lambda\|^2_S$ defined in (4.5.8) for representing the energy of the residual supported by the subset $\mathcal{S}$.

Our first result is concerned with the accuracy of approximation of the quantities $e_\Lambda(\mathcal{S})$ by their counterparts $e^\varepsilon_\Lambda(\mathcal{S})$. We have the following lemma

Lemma 4.7.4

Let Λ be a lower set, $\mathbf{u}^\varepsilon_\Lambda$ an approximate to $\mathbf{u}_\Lambda$ as in (4.7.21) and the quantities $e^\varepsilon_\Lambda(\mathcal{S})$ as above. Then we have for any $\mathcal{S} \subset \mathcal{M}(\Lambda)$

$$\sqrt{e_\Lambda(\mathcal{S})} - \sqrt{e^\varepsilon_\Lambda(\mathcal{S})} \le \|\mathbf{A}\|_S\|\mathbf{u}_\Lambda - \mathbf{u}^\varepsilon_\Lambda\| \le \varepsilon\|\mathbf{A}\|_S \tag{4.7.25}$$

Proof: In view of the the formulas (4.3.7) and (4.7.23), the function $(w_{\Lambda,\nu} - w^\varepsilon_{\Lambda,\nu}) \in V$ for $\nu \in \mathcal{M}$ is the unique solution of the system

$$-\Delta w = \text{ div } (\phi_\nu - \phi^\varepsilon_\nu) \quad \text{in} \quad D, \quad w_{|\partial D} = 0.$$

Therefore

$$\int_D |\nabla(w_{\Lambda,\nu} - w^\varepsilon_{\Lambda,\nu})|^2 = -\int_D (\phi_\nu - \phi^\varepsilon_\nu)\nabla(w_{\Lambda,\nu} - w^\varepsilon_{\Lambda,\nu})$$

We have seen in the proof of Lemma 4.3.1 that using only the form of the matrix $\mathbf{A}$ given in Lemma 4.2.1, one has for any $\mathbf{v} \in \ell^2(\mathcal{F}, V)$

$$\langle \mathbf{A}\mathbf{u}_\Lambda, \mathbf{P}_\mathcal{M}\mathbf{v}\rangle = \sum_{\nu\in\mathcal{M}} \int_D \phi_\nu \nabla \mathbf{v}_\nu,$$

This holds true if one replaces $\mathbf{u}_\Lambda$ by $\mathbf{u}^\varepsilon_\Lambda$ and the functions ϕ_ν by ϕ^ε_ν. We deduce then that

$$\|\mathbf{w} - \mathbf{w}^\varepsilon\|^2_{\ell^2(\mathcal{F},V)} = \sum_{\nu\in\mathcal{M}} \int_D |\nabla(w_{\Lambda,\nu} - w^\varepsilon_{\Lambda,\nu})|^2 = -\langle \mathbf{A}(\mathbf{u}_\Lambda - \mathbf{u}^\varepsilon_\Lambda), \mathbf{w} - \mathbf{w}^\varepsilon\rangle,$$

where $\mathbf{w}$ and $\mathbf{w}^\varepsilon$ are the vectors supported in $\mathcal{M}$ that have coordinates $w_{\Lambda,\nu}$ and $w^\varepsilon_{\Lambda,\nu}$ respectively. Since

$$|\langle \mathbf{A}(\mathbf{u}_\Lambda - \mathbf{u}^\varepsilon_\Lambda), \mathbf{w} - \mathbf{w}^\varepsilon\rangle| \le \|\mathbf{A}\|_S\|\mathbf{u}_\Lambda - \mathbf{u}^\varepsilon_\Lambda\| \ \|\mathbf{w} - \mathbf{w}^\varepsilon\|,$$

then $\|\mathbf{w} - \mathbf{w}^\varepsilon\| \le \|\mathbf{A}\|_S\|\mathbf{u}_\Lambda - \mathbf{u}^\varepsilon_\Lambda\|$ and for any $\mathcal{S} \subset \mathcal{M}$ it holds also $\|\mathbf{P}_\mathcal{S}(\mathbf{w} - \mathbf{w}^\varepsilon)\| \le \|\mathbf{A}\|_S\|\mathbf{u}_\Lambda - \mathbf{u}^\varepsilon_\Lambda\|$. Using the reverse triangular inequality, we deduce that

$$\|\mathbf{P}_\mathcal{S}\mathbf{w}\| - \|\mathbf{P}_\mathcal{S}\mathbf{w}^\varepsilon\| \le \|\mathbf{A}\|_S\|\mathbf{u}_\Lambda - \mathbf{u}^\varepsilon_\Lambda\|,$$

which is exactly the wanted result. ■

Given the previous result, we are able to get a bulk inequality of type (4.3.10) for the real residual knowing the approximate function $w^{\varepsilon}_{\Lambda,\nu}$. Indeed, it is easily checked that

$$\sqrt{e^{\varepsilon}_{\Lambda}(\mathcal{S})} \geq \theta \sqrt{e^{\varepsilon}_{\Lambda}(\mathcal{M})} + \varepsilon \|\mathbf{A}\|_S (1+\theta) \quad \Longrightarrow \quad \sqrt{e_{\Lambda}(\mathcal{S})} \geq \theta \sqrt{e_{\Lambda}(\mathcal{M})}. \tag{4.7.26}$$

However one needs the assumption to be well defined, that is

$$\sqrt{e^{\varepsilon}_{\Lambda}(\mathcal{M})} \geq \theta \sqrt{e^{\varepsilon}_{\Lambda}(\mathcal{M})} + \varepsilon \|\mathbf{A}\|_S (1+\theta). \tag{4.7.27}$$

Also to make full profit of previous adaptive approaches, a control on the cardinality of the intermediate set $\mathcal{S}$ is needed. We propose an incremental strategy for performing such tasks. We consider the following algorithm.

Algorithm 4.7.5

Let Λ be a lower set, $\mathcal{M}$ the margin of Λ, $0 < \theta_1 < \theta_2 < 1$ and $\varepsilon > 0$. For $j = 0, 1, \cdots$ do the following:

- *set $\varepsilon_j = \varepsilon/2^j$, compute $\mathbf{u}_{\Lambda}^{\varepsilon_j}$ appoximating $\mathbf{u}_{\Lambda}$ as in (4.7.21) and the associated $e^{\varepsilon_j}_{\Lambda}(\mathcal{M})$;*
- *If $\sqrt{e^{\varepsilon_j}_{\Lambda}(\mathcal{M})} < \varepsilon_j \|\mathbf{A}\|_S \frac{2(1+\theta_1)}{\theta_2 - \theta_1}$ then go to step $j+1$;*
- *Else output the smallest lower set $\mathcal{S}$ in $\mathcal{M}$ such that*

$$\sqrt{e^{\varepsilon_j}_{\Lambda}(\mathcal{S})} \geq \theta_1 \sqrt{e^{\varepsilon_j}_{\Lambda}(\mathcal{M})} + \varepsilon_j \|\mathbf{A}\|_S (1+\theta_1). \tag{4.7.28}$$

The previous algorithm always terminates and the relation (4.7.28) is well defined. Indeed, we have that $\varepsilon_j \to 0$ as j grows while $e^{\varepsilon_j}_{\Lambda}(\mathcal{M})$ becomes closer to $e_{\Lambda}(\mathcal{M}) > 0$, therefore the loop in j terminates. Moreover, when the loop terminates, we obtain $e^{\varepsilon_j}_{\Lambda}(\mathcal{M}) \geq \varepsilon_j \|\mathbf{A}\|_S \frac{1+\theta_1}{\theta_2 - \theta_1}$, hence

$$\sqrt{e^{\varepsilon_j}_{\Lambda}(\mathcal{M})} \geq \theta_2 \sqrt{e^{\varepsilon_j}_{\Lambda}(\mathcal{M})} \geq \theta_1 \sqrt{e^{\varepsilon_j}_{\Lambda}(\mathcal{M})} + 2\varepsilon_j \|\mathbf{A}\|_S (1+\theta_1) \geq \theta_1 \sqrt{e^{\varepsilon_j}_{\Lambda}(\mathcal{M})} + \varepsilon_j \|\mathbf{A}\|_S (1+\theta_1), \tag{4.7.29}$$

which justifies the existence of $\mathcal{S}$. Now we have the following Lemma

Lemma 4.7.6

Let Λ a lower set, $\mathcal{M}$ the margin of Λ and $0 < \theta_1 < \theta_2 < \min(\kappa(\mathbf{A}), \kappa(\mathbf{A})^{-\frac{1}{2}})$. Given

$\mathcal{S}$ output by Algorithm 4.7.5 and $\tilde{\Lambda} = \Lambda \cup \mathcal{S}$, one has

$$\|\mathbf{u} - \mathbf{u}_{\tilde{\Lambda}}\|_E \leq \delta_1 \|\mathbf{u} - \mathbf{u}_{\Lambda}\|_E, \tag{4.7.30}$$

where $\delta_1 = \sqrt{1 - \frac{\theta_1^2}{\kappa(\mathbf{A})}}$, and

$$\#(\mathcal{S}) \leq C_{\theta_2} \|(c_\nu)\|_{\ell_p^m(\mathcal{F})}^{1/s^*} \|\mathbf{r}_\Lambda\|^{-1/s^*}, \tag{4.7.31}$$

with the same expression for the constant C_{θ_2} used in Lemma 4.3.4.

Proof: Since (4.7.28) holds, then the implication (4.7.26) shows that $\sqrt{e_\Lambda(\mathcal{S})} \geq \theta \sqrt{e_\Lambda(\mathcal{M})}$, which by the reduction Lemma 4.3.3 implies the first inequality. As for the second inequality, we use the proof of Lemma 4.3.5 with the value θ_2 instead of θ. Choosing $\hat{\Lambda}$ as in there, we obtain that the set $\hat{\mathcal{S}} = \hat{\Lambda} \cap \mathcal{M}$ is lower in $\mathcal{M}$ and satisfies

$$\sqrt{e_\Lambda(\hat{\mathcal{S}})} \geq \theta_2 \sqrt{e_\Lambda(\mathcal{M})}, \quad \text{and} \quad \#(\hat{\mathcal{S}}) \leq C_{\theta_2} \|(c_\nu)\|_{\ell_p^m(\mathcal{F})}^{1/s^*} \|\mathbf{r}_\Lambda\|^{-1/s^*}.$$

We have then that

$$\sqrt{e_\Lambda^\varepsilon(\hat{\mathcal{S}})} \geq \theta_2 \sqrt{e_\Lambda^\varepsilon(\mathcal{M})} - \varepsilon_j \|\mathbf{A}\|_S (1+\theta_1) \geq \theta_1 \sqrt{e_\Lambda^{\varepsilon_j}(\mathcal{M})} + \varepsilon_j \|\mathbf{A}\|_S (1+\theta_1),$$

where we have used the over capturing of the bulk (4.7.29) in the last inequality. Since $\mathcal{S}$ is the smallest set with the property (4.7.28), then $\#(\mathcal{S}) \leq \#(\hat{\mathcal{S}})$ and the proof is complete. ∎

In view of the the previous lemma, we now are able to propose an adaptive algorithm that take into account the error on the approximation of the Galerkin projection. We assume we have $0 < \theta_1 < \theta_2 < \min(\kappa(\mathbf{A}), \kappa(\mathbf{A})^{-\frac{1}{2}})$, and we introduce the notation

$$\mathcal{S} := \text{OVERBULK}(\Lambda, \theta_1, \theta_2), \tag{4.7.32}$$

for the set output by Algorithm 4.7.5. we then consider the following

Algorithm 4.7.7

Define $\Lambda_1 := \{0_\mathcal{F}\}$ and compute $\mathbf{u}_{\Lambda_1} = u(0)$. For $n = 1, \cdots$ do the following:

- *Given that Λ_n has been defined, output $\mathcal{S}_n = \text{OVERBULK}(\Lambda_n, \theta_1, \theta_2)$.*
- *Enrich Λ_n by adding the element of $\mathcal{S}_n$, i.e. $\Lambda_{n+1} := \Lambda_n \cup \mathcal{S}_n$.*
- *Go to step $n+1$.*

By the exact same arguments of the previous section, it is easily proven that the previous algorithm yields index sets which are near optimal in the sense of (4.1.10).

4.8 Convergence of Galerkin approximation in the uniform sense

In this section, we investigate the convergence of the Galerkin approximations in the uniform sense. More precisely, given Λ a finite lower set of indices, we study how the Galerkin projection u_Λ can approximate the elements of the manifold $\mathcal{M} = \{u(y) : y \in U\}$ in the uniform sense, by inspecting the quantity

$$\|u - u_\Lambda\|_{\mathcal{V}_\infty} = \sup_{y \in U} \|u(y) - u_\Lambda(y)\|_V \tag{4.8.1}$$

The key point of our analysis is the following results in which we examine the stability of the uniform norm $\|\cdot\|_{\mathcal{V}_\infty}$ with respect the least square norm $\|\cdot\|_{\mathcal{V}_2}$ over spaces $\mathbb{V}_\Lambda$ for Λ lower.

Lemma 4.8.1

Let $\Lambda \subset \mathcal{F}$ be lower set. We have

$$\sup_{v \in \mathbb{V}_\Lambda \setminus \{0\}} \frac{\|v\|_{\mathcal{V}_\infty}}{\|v\|_{\mathcal{V}_2}} \leq \#(\Lambda). \tag{4.8.2}$$

Proof: Let $v = \sum_{\nu \in \Lambda} \mathbf{v}_\nu L_\nu$ be in $\mathbb{V}_\Lambda$. We have

$$\Big\| \sum_{\nu \in \Lambda} \mathbf{v}_\nu L_\nu \Big\|_{\mathcal{V}_\infty} \leq \sum_{\nu \in \Lambda} \|\mathbf{v}_\nu\|_V \|L_\nu\|_{L^\infty(U)} \leq \Big(\sum_{\nu \in \Lambda} \|\mathbf{v}_\nu\|_V^2 \Big)^{\frac{1}{2}} \Big(\sum_{\nu \in \Lambda} \|L_\nu\|_{L^\infty(U)}^2 \Big)^{\frac{1}{2}} = \sqrt{K_{0,0}(\Lambda)} \|v\|_{\mathcal{V}_2},$$

where we have defined

$$K_{0,0}(\Lambda) := \sum_{\nu \in \Lambda} \|L_\nu\|_{L^\infty(U)}^2 = \sum_{\nu \in \Lambda} \prod_{j \geq 1} (2\nu_j + 1). \tag{4.8.3}$$

This definition coincides with the definition (A.4.6) given in the appendix. We show there, Lemma A.4.1, that for Λ lower, one has $K_{0,0}(\Lambda) \leq (\#(\Lambda))^2$. The proof is then complete. ∎

The Legendre polynomials L_ν attain all there supremums over U on the point $(1, 1, 1, \ldots)$, therefore by setting $\mathbf{v}_\nu = \|L_\nu\|_{L^\infty(U)}$ in the previous proof, we get only equalities, showing that the supremum of the ratio of the norms is actually equal to $\sqrt{K_{0,0}(\Lambda)}$. We should also mention that $K_{0,0}(\Lambda)$ is equal to $(\#(\Lambda))^2$ for Λ of rectangular block shape, see Appendix. However, for anisotropic lower sets Λ, the bound $(\#(\Lambda))^2$ might be overestimated. Also, let us remark that the result of the lemma is valid regardless the shape of Λ.

The stability result is not usable immediately for relating (4.8.1) to the least square norm of $u - u_\Lambda$ since the latter does not belong to a polynomial space. Instead, we have the following result

Lemma 4.8.2

Let Λ by a lower set of cardinality n and $v \in \mathbb{V}_\Lambda$, we have

$$\|u - v\|_{\mathcal{V}_\infty} \leq 3n\|u - v\|_{\mathcal{V}_2} + 4\left\|(\|u_\nu\|_V)\right\|_{\ell^p_m(\mathcal{F})}(n+1)^{1-s^*}, \quad s^* = \frac{1}{2} - \frac{1}{p} \tag{4.8.4}$$

Proof: Let Λ_n^{L*} and Λ_n^{P*} the lower sets used in the best n-term approximation (4.1.10) and (4.1.11). We set $\Lambda' = \Lambda \cup \Lambda_n^{L*} \cap \Lambda_n^{P*}$. We have that Λ' is lower as the union of three lower sets and is cordiality smaller than $3n$. We have for any $w \in V_{\Lambda'}$ that $v - w \in V_{\Lambda'}$, hence

$$\|u-v\|_{\mathcal{V}_\infty} \leq \|u-w\|_{\mathcal{V}_\infty} + \|w-v\|_{\mathcal{V}_\infty} \leq \|u-w\|_{\mathcal{V}_\infty} + 3n\|w-v\|_{\mathcal{V}_2} \leq \|u-w\|_{\mathcal{V}_\infty} + 3n\|u-w\|_{\mathcal{V}_2} + 3n\|u-v\|_{\mathcal{V}_2}$$

hence

$$\|u - v\|_{\mathcal{V}_\infty} \leq 3n\|u - v\|_{\mathcal{V}_2} + \inf_{w \in V_{\Lambda'}} \left(\|u - w\|_{\mathcal{V}_\infty} + 3n\|u - w\|_{\mathcal{V}_2}\right)$$

The proof is finished by taking w to be the Legednre series of u, i.e $w = \sum_{\nu \in \Lambda} v_\nu L_\nu$, using (4.1.10) and (4.1.11) and the property $\|v_\nu\|_V \leq \|u_\nu\|_V$ for any $\nu \in \mathcal{F}$ ■

The previous lemma show that if v is any approximation that is near optimal in the sense of (4.1.10), then it also yields convergence rate $1 - s^* = \frac{1}{2} - s = \frac{3}{2} - p$ in the uniform sense, which shows that only a deterioration $(n+1)^{\frac{1}{2}}$ is to be expected in the uniform sense when comparing with the best benchmark (4.1.10). In particular, for small values of p, namely $p < \frac{3}{2}$, the Legendre series associated with the sets Λ_n^{L*} provide convergence to u simultaneously in the uniform and least square sense.

4.9 Conclusion

In this chapter, we have introduced adaptive algorithms for sparse Galerkin approximation of the solution map u of the elliptic model in the mean squares sense. These algorithms have remarkable features, in particular:

(i) They builds the Galerkin projection by an iterative method which at every step amounts to solving a number of boundary value problems with fixed stiffness matrix as in Chapter 3.

(ii) From a theoretical point of view, the convergence with respect to the polynomial dimension can be proved to be near optimal in the mean square sense and slightly deteriorated in the uniform sense.

The analysis is strongly tied to the linearity of the model and the affine dependence on y. It adaptation to other models might not be possible. For example, it is not obvious how to adapt the analysis for the semi-linear problem

$$u^3 - \operatorname{div}(a\nabla u) = f \tag{4.9.1}$$

where a is still affine in y.

The remainder of the thesis is concerned with methods that can be applied for more general PDEs. We investigate in particular collocation methods, in others words methods that are based only on the instances of the solution map u. We present an interpolation scheme and a least square scheme and analyze theirs convergence and theirs numerical effectiveness.

Part III

Non-intrusive adaptive algorithms

Chapter 5

Sparse high-dimensional polynomial interpolation

Contents

5.1 Introduction

In this chapter, we introduce an interpolation scheme which can be used for non-intrusive treatment of parametric PDE. This scheme gathers different aspects of poly-

nomial approximation in high dimensions. In particular, it can be seen as a collocation method as in [4, 7, 8], as a sparse grid method as in [9, 50, 70, 69] or as a sparse polynomial approach as in [34, 33, 22, 53, 25].

We recall that the setting we are interested in is the setting of parametric PDEs of the form

$$\mathcal{D}(u, y) = 0, \tag{5.1.1}$$

where $u \mapsto \mathcal{D}(u, y)$ is a partial differential operator that depends on an infinite number of parameters y_j represented by the parameter vector $y = (y_j)_{j\geq 1} \in U := [-1, 1]^{\mathbb{N}}$. We assume that the problem (5.1.1) is well posed in some Banach space V for any y, so that we may defined the solution map u by

$$y \in U \mapsto u(y) \in V. \tag{5.1.2}$$

We have seen in chapters 1-2 that under a mild anisotropic dependence of the parametric PDE on the parameter y, the solution map u can be approximated by multi-variate polynomials in y with algebraic rates. For the typical example of parametric elliptic PDEs studied in Chapter 1, given by the equation (1.1.1) with affine parameter dependence (1.1.2) and uniform ellipticity assumption $\mathbf{UEA}(r, R)$ (1.1.3), it is proven that if $(\|\psi_j\|_L)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$, then the solution maps u can be approximated by its Taylor series with convergence rate $(n + 1)^{-s}$ with $s = \frac{1}{p} - 1$. More generally, we have seen in Chapter 2 that when the operator $\mathcal{D}$ depends on y through the expansion $\sum_{j\geq 1} \psi_j y_j$ where the ψ_j are functions in some Banach space L, then under the assumptions of Theorem 2.4.3, in which the anisotropy assumption $(\|\psi_j\|_L)_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $p < 1$ is crucial, then the solution map u can be approximated by its Legendre series with convergence rate $(n+1)^{-s}$, $s = \frac{1}{p} - 1$ in the uniform sense and convergence rate $(n + 1)^{-s^*}$, $s^* = \frac{1}{p} - \frac{1}{2}$ in the mean square sense.

We use the notations of the set of multi-indices $\mathcal{F}$, the Legendre polynomials $(L_\nu)_{\nu\in\mathcal{F}}$, the spaces $\mathcal{V}_\infty$, $\mathcal{V}_2$ and theirs norms is as in §1.2 of Chapter 1 and define the polynomials spaces

$$\mathbb{V}_\Lambda := V \otimes \mathbb{P}_\Lambda, \qquad \mathbb{P}_\Lambda := \operatorname{span}\{L_\nu \ : \ \nu \in \Lambda\}. \tag{5.1.3}$$

for $\Lambda \subset \mathcal{F}$. Under the assumption of Theorem 2.4.3, there exists a sequence $(\Lambda_n)_{n\geq 1}$ of nested sets with $\Lambda_n = n$ for which we have an approximation in the uniform sense, i.e.

$$\inf_{v\in\mathbb{V}_{\Lambda_n}} \|u - v\|_{\mathcal{V}_\infty} \leq C(n + 1)^{-s}, \qquad s := \frac{1}{p} - 1, \tag{5.1.4}$$

and an other sequence $(\Lambda_n)_{n\geq 1}$ of nested sets with $\Lambda_n = n$ for which we have an approximation in the mean square sense,

$$\inf_{v\in V_{\Lambda_n}} \|u - v\|_{\mathcal{V}_2} \leq C'(n + 1)^{-s^*}, \qquad s^* := \frac{1}{p} - \frac{1}{2}, \tag{5.1.5}$$

The previous rates where obtained through approximations by truncated Legendre series, Theorem 2.2.2 and implications. We also recall that the sets Λ_n can be chosen lower, i.e.

$$\nu \in \Lambda_n \text{ and } \mu \leq \nu \Rightarrow \mu \in \Lambda_n, \tag{5.1.6}$$

The previous results are purely theoretical, we have not discussed for the abstract equation (5.1.1) what are the strategies for computable approximations that preserve the convergence rates. However we have seen, for the elliptic model (1.1.1) of Chapter 1 with the assumption $(\|\psi_j\|_{L^\infty(D)})_{j\geq 1} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$, that in practice:

(i) Taylor expansions (Chapter 3) associated with lower sets can be recursively computed. Adaptive methods based on such expansions have been proved to converge in the uniform sense with the same rate as in (5.1.4).

(ii) Projection methods (Chapter 4) can be built adaptively using techniques of a-posteriori analysis. We have proved the convergence in the mean square sense with the same rate as in (5.1.5) and with a rate $s - \frac{1}{2}$ in the uniform sense. Galerkin projections are also considered and analyzed in [1, 5, 7, 34, 56, 57]

For the elliptic model (1.1.1) of Chapter 1, collocation methods [4, 7, 8, 66, 70, 69] in the setting $d < \infty$ can also produce polynomial approximations that converge toward u with prescribed rates. Such approximations $u_{\Lambda_k} \in V_{\Lambda_k}$ are only based on particular solution instances $u(y^i)$ at well chosen values $y^1, \ldots, y^{\#(\Lambda_k)} \in U$ of the parameter vector.

In contrast with the two first approaches, one significant advantage of the last approach is that it is non intrusive and can then be applied for general parametric PDE of the form(5.1.1). The instances $u(y^i)$ are obtained by a numerical solver for the problem (5.1.1) then polynomial approximations are built from these instances by numerical techniques similar to those employed for scalar valued maps such as, sparse grids, interpolation or least squares. However, the theoretical analysis of collocation methods is less satisfactory in the sense that convergence rates similar to (5.1.4) and (5.1.5) do not seem to have been established. This is in part due to the rigidity of the index sets Λ_k that are considered, which grow rapidly in cardinality and the difficulty to control the stability of interpolation operators in arbitrary high dimension. In addition, adaptive methods for building the sets Λ_k have not been much developed in the collocation framework. We ask then the following legitimate questions:

- *Given a parametric PDE such that the solution map can be approximated as in* (5.1.4) *and* (5.1.5)*, can one obtain easily approximations* $u_{\Lambda_n} \in V_{\Lambda_n}$ *using collocation methods and that converge with the same rates?*

- *Since* Λ_{n+1} *is in general obtained from* Λ_n *by an enrichment procedure, is it possible to compute* $u_{\Lambda_{n+1}}$ *easily using* u_{Λ_n} *?*

- *How to control the stability of the collocation methods?*

The objective of this chapter is to propose and study a collocation method that satisfies these three prescriptions. The method is based on a high dimensional interpolation process that can naturally be coupled with an adaptive selection of the polynomial spaces. We construct an interpolation operator $\mathcal{I}_\Lambda$ that maps real or complex valued functions defined on U into $\mathbb{P}_\Lambda$. A standard vectorization yields an interpolation operator that maps V-valued functions defined on U into V_Λ with the same properties. Namely, given Λ_n with $\#(\Lambda_n) = n$, we compute an approximation $\mathcal{I}_{\Lambda_n} u \in V_{\Lambda_n}$ that coincides with u at n well chosen point $y^1, \ldots, y^n$ of U. We do not address the questions of unisolvency between the interpolation points and polynomials space with which references such as [58, 60, 36] are concerned. Our approach consist in generalizing the sparse grids interpolation methods [76, 50, 71, 7, 8, 70, 69], considered with specific polynomials space , to more general polynomial spaces V_Λ with Λ being any lower set.

This chapter is organized as follows. In §5.2.1, we build for any lower set, a grid of points Γ_Λ which is unisolvent for $\mathbb{P}_\Lambda$ and provide a definition of the interpolation operator $\mathcal{I}_\Lambda$. This construction is based on a univariate sequence of points $(z_k)_{k\geq 0}$ and a standard tensorization and sparsification technique, originally due to Smolyak [76] and that is already addressed for many type of lower sets, see for example [7, 8, 70, 69]. The main feature of this process is the inherent nested structures the grids, which is well adapted to an adaptive construction of the index set: the enrichment of Λ by one index is reflected by the enrichment of Γ_Λ by one point. The amount of computation is therefore minimized since all previously computed solution instances are used.

In §5.2.2, we establish a simple formula for the computation of the increments $\mathcal{I}_{\Lambda_{n+1}} u - \mathcal{I}_{\Lambda_n} u$ where $\Lambda_n \subset \Lambda_{n+1}$ are both lower and differ only by one index. We show that the computation of such increments can be done by a Newton like formula in one dimension that only requires, beside the evaluation of u in the new point $\Gamma_{\Lambda_{n+1}} \backslash \Gamma_{\Lambda_n}$, at most $2\#(\Lambda_n)$ usual operations in V. We shall show in particular that the incremental procedure can be used to compute $\mathcal{I}_\Lambda$ for any Λ lower at the cost of $\#(\Lambda)$ evaluations of u plus at most $(\#(\Lambda))^2$ product and sum operation in V.

In §5.3, we study the stability of the interpolation operator $\mathcal{I}_\Lambda$. In particular we establish bounds on the Lebesgue constant which only depends on the cardinality of the set Λ (not on its shape or on the parametric dimension, here infinite $d = \infty$). These bounds grow algebraically with $\#(\Lambda)$, provided that bounds that grow algebraically with $(k+1)$ are available for the Lebesgue constants associated to the sections $\{z_0, \ldots, z_k\}$ of the sequence $(z_k)_{k\geq 0}$. Algebraic growth are available in the univariate case for the so-called $\Re$-Leja point [18, 19, 21] and will be recalled in Chapter 6.

When combining the approximation estimate (5.1.4) together with a bound of the form $(\#(\Lambda_n))^b$ for the Lebesgue constant associated with $\mathcal{I}_{\Lambda_n} u$, one expects the interpolation $\mathcal{I}_{\Lambda_n} u$ to converge towards u with a deteriorated rate $(n+1)^{-(s-b)}$. In §5.4, we show by a different error analysis that, under similar assumptions as in Chapter 2, one

can construct a sequence $(\Lambda_n)_{n\geq 0}$ for which the approximations $\mathcal{I}_{\Lambda_n}u$ converge towards u with the optimal rate $(n+1)^{-s}$. We present in §5.5 numerical results that illustrate the performance of our adaptive interpolation scheme.

5.2 Interpolation on nested grids

As we already mentioned, we provide in this section the construction of the interpolation operator associated with a lower set and corresponding unisolvent grid, following the technique of Smolyak [76]. Here, we consider scalar valued functions and work in the infinite dimension setting $d = \infty$.

5.2.1 The sparse interpolation operator

Let $(z_k)_{k\geq 0}$ be any sequence of pairwise distinct points in $[-1,1]$. We denote by I_k the univariate Lagrange polynomial interpolation operators associated with the sections $\{z_0,\ldots,z_k\}$. Theses operators act on functions $g \in C([-1,1])$ according to $I_0 g = g(z_0)\mathbf{1}$ and for $k \geq 1$

$$I_k g := \sum_{i=0}^{k} g(z_i) l_i^k, \quad \text{where} \quad l_i^k(t) := \prod_{\substack{i=0\\ i\neq j}}^{k} \frac{t - z_i}{z_j - z_i}. \tag{5.2.1}$$

The polynomials $l_0^k,\ldots,l_k^k$ are the Lagrange polynomials associated with $\{z_0,\ldots,z_k\}$. The operator I_k is a projection operator on the space $\mathbb{P}_k$ of univariate polynomials of degree at most k. We introduce the difference operators

$$\Delta_k := I_k - I_{k-1}, \qquad k \geq 0, \tag{5.2.2}$$

with the convention that I_{-1} is the null operator. The operator Δ_k is an increment operator which is used to update the operator I_{k-1} into I_k. Now given $\nu = (\nu_1, \nu_2, \ldots) \in \mathcal{F}$, we define tensorized operators $I_\nu = \otimes_{j\geq 1} I_{\nu_j}$ and $\Delta_\nu = \otimes_{j\geq 1}\Delta_{\nu_j}$ on the space $C(U)$ of continuous functions over U by the following: For the null multi-index $\nu = 0_{\mathcal{F}}$, one has

$$\Delta_{0_{\mathcal{F}}} g = I_{0_{\mathcal{F}}} g = g(z_{0_{\mathcal{F}}})\mathbf{1}, \qquad z_{0_{\mathcal{F}}} := (z_0, z_0, \ldots) \in U, \tag{5.2.3}$$

then for a given multi-index $\nu \neq 0_{\mathcal{F}}$ supported in $\{1,\ldots,J\}$, in other words $\nu_j = 0$ for $j > J$, one has

$$I_\nu g := (\otimes_{j=1}^{J} I_{\nu_j}) g_J \quad \text{and} \quad \Delta_\nu g := (\otimes_{j=1}^{J} \Delta_{\nu_j}) g_J, \tag{5.2.4}$$

where g_J is the function defined over $[-1,1]^J$ by

$$g_J(y_1,\ldots,y_J) := g(y_1,\ldots,y_J,z_0,z_0,\ldots) \tag{5.2.5}$$

and the tensorizations $\otimes_{j=1}^{J}$ are defined in the usual sense.

For $\Lambda \subset \mathcal{F}$ an index set, we associate the operator $\mathcal{I}_\Lambda$ and the grid Γ_Λ defined by

$$\mathcal{I}_\Lambda := \sum_{\nu \in \Lambda} \Delta_\nu, \qquad \Gamma_\Lambda := \Big\{ z_\nu := (z_{\nu_1}, z_{\nu_2}, \ldots) \,:\, \nu \in \Lambda \Big\}. \tag{5.2.6}$$

Thanks to the tensor product structure of the rectangular blocks $\mathcal{B}_\nu := \{\mu \leq \nu\}$ and telescopic cancellations, we have that

$$\mathcal{I}_{\mathcal{B}_\nu} = \sum_{\mu \leq \nu} \otimes_{j \geq 1} \Delta_{\mu_j} = \otimes_{j \geq 1} \sum_{\mu_j \leq \nu_j} \Delta_{\mu_j} = \otimes_{j \geq 1} I_{\nu_j} = I_\nu, \tag{5.2.7}$$

is effectively the interpolation operator for the tensor product polynomial space $\mathbb{P}_{\mathcal{B}_\nu}$ associated with the tensor grid $\Gamma_{\mathcal{B}_\nu}$. A similar result holds for lower sets Λ in general. The lower sets Λ might significantly differ from the sparse grid sets which are usually considered in the literature [76, 9, 50, 7, 8, 70, 69]. However, the argument showing that $\mathcal{I}_\Lambda$ is the polynomial interpolation operator on $\mathbb{P}_\Lambda$ associated with the grid Γ_Λ is very similar. For convenience of the reader, we give a precise statement of this result.

Theorem 5.2.1

For $\Lambda \subset \mathcal{F}$ lower, the grid Γ_Λ is unisolvent for $\mathbb{P}_\Lambda$ and for any function g defined on U, the unique element in $\mathbb{P}_\Lambda$ which agrees with g on Γ_Λ is given by $\mathcal{I}_\Lambda g$.

Proof: We consider $\nu \in \Lambda$ and notice that the lower structure of Λ implies that $\mathcal{B}_\nu \subset \Lambda$. In view of (5.2.6) and the observation that $\mathcal{I}_{\mathcal{B}_\nu}$ is the interpolation operator associated with $\Gamma_{\mathcal{B}_\nu}$ to which z_ν belongs, we infer

$$\mathcal{I}_\Lambda g(z_\nu) = \mathcal{I}_{\mathcal{B}_\nu} g(z_\nu) + \sum_{\mu \in \Lambda : \mu \not\leq \nu} \Delta_\mu g(z_\nu) = g(z_\nu) + \sum_{\mu \in \Lambda : \mu \not\leq \nu} \Delta_\mu g(z_\nu).$$

Given $\mu \not\leq \nu$, there exists at least one j such that $\nu_j < \mu_j$. Since the points of the univariate interpolations are nested, then the univariate operator Δ_{μ_j} returns a polynomial which vanishes at z_{ν_j}, and so $\Delta_\mu g$ vanishes at all points with j^{th} coordinate equal to z_{ν_j}. In particular, $\Delta_\mu g(z_\nu) = 0$ and we have thus proved that $\mathcal{I}_\Lambda g(z_\nu) = g(z_\nu)$ for any $\nu \in \mathcal{F}$. ■

The fact that Γ_Λ is unisolvent for the polynomial space $\mathbb{P}_\Lambda$ when Λ is lower appears to be known from early works on polynomial interpolation, see Chapter IV in the book [58] in which bivariate polynomials associated to lower sets are referred to as "polynômes pleins". This also appears as a particular case of the theory of the "least polynomial spaces" for interpolation of functions on general multivariate point sets, see in particular [36]. In the previous reference, polynomials associated to lower sets Λ are referred to as "order closed polynomials" and the spaces they generate are proved to be the least polynomial spaces for sets of the form Γ_Λ.

In order to establish the unisolvency, we have not used particularly that polynomials defined over U are involved. One can generalize the construction in a straightforward

way to tensorized domains of the more general form $U = \prod_{j\geq 1} U_j$ with a different univariate sequence $(z_k^j)_{k\geq 0}$ in each coordinate domain U_j. Another straightforward generalization is when the univariate polynomial spaces $\mathbb{P}_k$ are replaced by more general nested spaces S_k such that $\{z_0, \ldots, z_k\}$ is unisolvent for S_k. In such case, Γ_Λ is unisolvent for the space

$$S_\Lambda = \oplus_{\nu\in\Lambda}(\otimes_{j\geq 1} S_{\nu_j}), \tag{5.2.8}$$

which generalizes $\mathbb{P}_\Lambda$ and the interpolation operator is defined in a similar manner as $\mathcal{I}_\Lambda$. Sparse grid interpolation based on hierarchical finite element spaces are a particular instance of this generalization.

The previous construction can also be generalized in order to meet the framework of isotropic and anisotropic sparse grids as discussed in [7, 8, 70, 69]. We recall in particular the notations (3.29) and (3.30) introduced in the general introduction for a unified description of such methods. We denote by m a strictly increasing function from $\mathbb{N}$ into $\mathbb{N}$ that satisfies $m(0) = 0$ and the convention $m(-1) = -1$. Given a lower set Λ, we introduce the notation

$$m(\Lambda) := \bigcup_{i\in\Lambda} B_{m(i)} \quad \text{with} \quad B_{m(i)} := \Big\{\nu \in \mathcal{F} : m(i_j - 1) < \nu_j \leq m(i_j), j \geq 1\Big\}. \tag{5.2.9}$$

The set $m(\Lambda)$ is a union of $\#(\Lambda)$ adjacent blocks, that coincides with Λ if m is the identity function, and it is also a lower set. It resembles in shape to Λ if m is a doubling rule, i.e. $m(k) = 2^k$ for $k \geq 1$. We remark that a direction j is active in $m(\Lambda)$ if and only if it is active in Λ. In particular Λ and $m(\Lambda)$ are supported in the same support. We now define, using Smolyak formula as in (3.30), the following operator

$$\mathcal{I}_{m(\Lambda)} := \sum_{i\in\Lambda} \otimes_{j\geq 1}(I_{m(i_j)} - I_{m(i_j-1)}). \tag{5.2.10}$$

The notation $\mathcal{I}_{m(\Lambda)}$ is justified by the fact the right hand is equal to the sum in (5.2.6). Indeed, thanks to the telescopic sum, we have for any $i \in \Lambda$ that

$$\otimes_{j\geq 1}(I_{m(i_j)} - I_{m(i_j-1)}) = \sum_{\nu\in B_{m(i)}} \Delta_\nu. \tag{5.2.11}$$

Therefore $\mathcal{I}_{m(\Lambda)}$ is an interpolation operator associated with the sparse grids of points $\Gamma_{m(\Lambda)}$ defined as in (5.2.6), with $m(\Lambda)$ instead of Λ. This generalize the results in [7, 8, 70, 69] in which this was established for many types of lower sets Λ.

We should also note an interesting property of the Smolyak construction. For Λ lower, the operator $\mathcal{I}_\Lambda$ is an interpolation operator, so that in particular it is a projection operator over $\mathbb{P}_\Lambda$. Another simple way to see this is by imitating the argument of the proof of Theorem 5.2.1, namely for $\nu \in \Lambda$ we have

$$\mathcal{I}_\Lambda y^\nu = I_\nu y^\nu + \sum_{\mu\in\Lambda:\mu\not\leq\nu} \Delta_\mu y^\nu = \prod_{j\geq 1} I_{\nu_j} y_j^{\nu_j} + \sum_{\mu\in\Lambda:\mu\not\leq\nu} \prod_{j\geq 1} \Delta_{\mu_j} y_j^{\nu_j} = y^\nu, \tag{5.2.12}$$

where we have used that $I_k y^k = y^k$ for any $k \geq 0$ and $\Delta_{\mu_j} y_j^{\nu_j} = I_{\mu_j} y_j^{\nu_j} - I_{\mu_j - 1} y_j^{\nu_j} = y_j^{\nu_j} - y_j^{\nu_j} = 0$, whenever $\nu_j < \mu_j$ for some j.

We remark that, for the previous result, we have only used the fact that for every $k \geq 0$, the operator I_k is a projection operator over the space $\mathbb{P}_k$. Therefore, the Smolyak construction also yield projection operators over $\mathbb{P}_\Lambda$ by tensorization of projection operators over $\mathbb{P}_k$. Having said that, we need to give a sense to infinite tensorization of projection operators.

If the operator I_0 is a projection from $C([-1,1])$ into $\mathbb{P}_0$ the space of constant polynomials, then it is defined by $I_0 g = p_0(g)\mathbf{1}$ where p_0 is linear form over $C([-1,1])$. Since $I_0 \mathbf{1} = \mathbf{1}$, then $p_0(\mathbf{1}) = 1$, so that necessarily p_0 is defined over $C([-1,1])$ by

$$p_0(g) = \int_{-1}^{1} g(t) d\lambda(t), \tag{5.2.13}$$

where λ a measure over $[-1,1]$ with total mass 1. The infinite dimensional tensorization can be obtained by considering the measure λ_∞ defined by $d\lambda_\infty(y) = \otimes_{j \geq 1} d\lambda(y_j)$ and defining $I_{0_\mathcal{F}}$ and the function g_J in (5.2.5) by

$$I_{0_\mathcal{F}} g := \int_U g(z) d\lambda_\infty(z), \qquad g_J(y_1, \ldots, y_J) := \int_U g(y_1, \ldots, y_J, z) d\lambda_\infty(z). \tag{5.2.14}$$

We have then the following lemma

Lemma 5.2.2

Let $I_{-1}, I_0, I_2, \cdots$ a family of operators such that $I_{-1} = 0$, $I_0 g = p_0(g)\mathbf{1}$ with p_0 as in (5.2.13) and every operator I_k is a projection operator over $\mathbb{P}_k$. For any $\Lambda \subset \mathcal{F}$ lower, the operator $\mathcal{I}_\Lambda$ defined as in (5.2.6) is a projection operator over $\mathbb{P}_\Lambda$.

A typical and interesting setting for the previous lemma is when the operators I_k are polynomial interpolation operators associated with sets of mutually disjoint points $\{z_0^k, \ldots, z_k^k\}$ which are not necessarily nested. For instance the simple roots of a family of orthogonal polynomials. The resulting operator $\mathcal{I}_\Lambda$ in this case also generalize a type of collocation methods on sparse grids [7, 8, 70, 69].

Finally, we should note that in the setting of the previous lemma, one can study the stability of the operator $\mathcal{I}_\Lambda$ through the study of Lebesgue constant, thanks to the following classical inequality always valid with projectors

$$\|g - \mathcal{I}_\Lambda g\| \leq (1 + \mathbb{L}_\Lambda) \inf_{P \in \mathbb{P}_\Lambda} \|g - P\|, \tag{5.2.15}$$

where $\|\cdot\|$ is any given norm over $C(U)$ and

$$\mathbb{L}_\Lambda := \sup_{P \in C(U) : \|P\| = 1} \|\mathcal{I}_\Lambda P\|. \tag{5.2.16}$$

The study of the growth of Lebesgue constants will be investigated in more details in §5.3.

5.2.2 A Newton like recursive formula

The interpolation process introduced in the previous section can be seen as a generalization of Lagrange polynomial interpolation on nested sets of points in dimension 1. In such setting, the hierarchical computation of the operators I_k is well understood. Indeed, one has the representation by finite Newton series, for $g \in C([-1,1])$ one has

$$I_k g = \sum_{j=0}^{k} [g(z_0), \dots, g(z_j)](z - z_j) \dots (z - z_j), \quad k \geq 0 \tag{5.2.17}$$

where $([g(z_0), \dots, g(z_j)])_{j\geq 0}$ are the so-called divided difference associated with g and the sequence $(z_j)_{j\geq 0}$, see [35] for more details. Up to a renormalization, the previous form is equivalent to expanding the additive polynomial increment $\Delta_k g = I_k g - I_{k-1} g$, which update the polynomial $I_{k-1} g$ into $I_k g$, in the basis of Lagrange polynomials associated with $\{z_0, \dots, z_k\}$ according to Lagrange interpolation formula. Since for every $j \leq k$, we have $I_k g(z_j) = I_{k-1} g(z_j) = g(z_j)$, then we have

$$\Delta_k g = \Big(g(z_k) - I_{k-1} g(z_k) \Big) h_k, \tag{5.2.18}$$

where the hierarchical polynomials $(h_k)_{k\geq 0}$ are defined by

$$h_0 := \mathbf{1}, \quad \text{and} \quad h_k(t) := \prod_{j=0}^{k-1} \frac{t - z_j}{z_k - z_j} \quad \text{for} \quad k \geq 1. \tag{5.2.19}$$

A similar result holds in the multi-variate setting. We define the tensorized hierarchical polynomials $(\mathbf{H}_\nu)_{\nu\in\mathcal{F}}$ defined by

$$\mathbf{H}_\nu(y) := \prod_{j\geq 1} h_{\nu_j}(y_j), \qquad y := (y_j)_{j\geq 1} \in U. \tag{5.2.20}$$

We have the following

Lemma 5.2.3

Let $\Lambda \subset \mathcal{F}$ be a lower set and ν an index in $\mathcal{F}$ such that $\Lambda' = \Lambda \cup \{\nu\}$ is also lower. Then, it holds

$$\Delta_\nu g = \mathcal{I}_{\Lambda'} g - \mathcal{I}_\Lambda g = g_\nu \mathbf{H}_\nu, \qquad g_\nu := g(z_\nu) - \mathcal{I}_\Lambda g(z_\nu). \tag{5.2.21}$$

Proof: On the one hand, in view of (5.2.18), the tensorization of the operators Δ_{ν_j} as described in (5.2.4) necessarily yields a difference operator satisfying

$$\Delta_\nu g = g_\nu \otimes_{j=1}^{J} h_{\nu_j} = g_\nu \mathbf{H}_\nu, \tag{5.2.22}$$

where J is an integer such that ν is supported in $\{1,\dots,J\}$ and g_ν is a constant depending on the values of the function g on the tensorized grid of interpolation points $\Gamma_{\mathcal{B}_\nu}$. On the other hand, similarly to the one dimensional case above, using Lagrange interpolation formula in order to expresses the polynomial $\Delta_\nu g \in \mathbb{P}_{\Lambda'}$ in the basis of Lagrange polynomials $(l_{\Lambda',\mu})_{\mu\in\Lambda'}$ associated with $\mathbb{P}_{\Lambda'}$ and $\Gamma_{\Lambda'}$, we obtain

$$\Delta_\nu g = \mathcal{I}_{\Lambda'} g - \mathcal{I}_\Lambda g = \sum_{\mu\in\Lambda'} \Big(\mathcal{I}_{\Lambda'} g(z_\mu) - \mathcal{I}_\Lambda g(z_\mu)\Big) l_{\Lambda',\mu} = \Big(g(z_\nu) - \mathcal{I}_\Lambda g(z_\nu)\Big) l_{\Lambda',\nu}, \tag{5.2.23}$$

where we have used that for $\mu \in \Lambda \subset \Lambda'$, one has $\mathcal{I}_{\Lambda'} g(z_\mu) = \mathcal{I}_\Lambda g(z_\mu) = g(z_\mu)$ and $\mathcal{I}_{\Lambda'} g(z_\nu) = g(z_\nu)$. Comparing (5.2.22) and (5.2.23) shows that $\mathbf{H}_\nu$ and $l_{\Lambda',\nu}$ are equal up to constant. Since $\mathbf{H}_\nu(z_\nu) = l_{\Lambda',\nu}(z_\nu) = 1$, then they are actually equal, which finishes the proof. ■

Let us remark that, in view of $\Delta_\mu g(z_\nu) = 0$ for $\mu \not\leqslant \nu$ showed in the proof of Theorem 5.2.1, the quantity g_ν defined in the previous lemma satisfies

$$g_\nu = g(z_\nu) - \sum_{\mu\in\Lambda:\mu<\nu} \Delta_\mu g(z_\nu) = g(z_\nu) - \sum_{\mu\in\Lambda:\mu<\nu} g_\mu \mathbf{H}_\mu(z_\nu), \tag{5.2.24}$$

depends effectively only on the grid of points $\Gamma_{\mathcal{B}_\nu}$.

As explained in the introduction, we are interested in performing polynomial interpolation for a nested sequence of lower sets $(\Lambda_n)_{n\geq 1}$ with $n = \#(\Lambda_n)$. Accordingly the grids $(\Gamma_{\Lambda_n})_{n\geq 1}$ are also nested. The sets Λ_n may either be fixed in advance, or adaptively chosen based on information gained at earlier computational steps. We have that for $n \geq 2$ that $\Lambda_n = \Lambda_{n-1} \cup \{\nu^n\}$ for some multi-index ν^n, therefore using the previous results, we have

$$\mathcal{I}_{\Lambda_n} g = \mathcal{I}_{\Lambda_{n-1}} g + \Big(g(z_{\nu^n}) - \mathcal{I}_{\Lambda_{n-1}} g(z_{\nu^n})\Big)\mathbf{H}_{\nu^n}. \tag{5.2.25}$$

We have then the following lemma.

Lemma 5.2.4

Let $(\Lambda_n)_{n\geq 1}$ be a sequence of nested lower sets with $n = \#(\Lambda_n)$ and denote by $(\nu^k)_{k\geq 1} \in \mathcal{F}^{\mathbb{N}}$ the indices such that $\Lambda_1 = \{\nu^1\}$ and for $n \geq 2$, $\Lambda_n = \Lambda_{n-1} \cup \{\nu^n\} = \{\nu^1,\dots,\nu^n\}$. For any $n \geq 1$, we have

$$I_{\Lambda_n} g = \sum_{k=0}^{n} g_{\nu^k} \mathbf{H}_{\nu^k} \tag{5.2.26}$$

where the coefficients g_{ν^k} are defined recursively by

$$g_{\nu^1} = g(z_{0_{\mathcal{F}}}), \qquad g_{\nu^k} := g(z_{\nu^k}) - \mathcal{I}_{\Lambda_{n-1}} g(z_{\nu^k}) = g(z_{\nu^k}) - \sum_{i=1}^{k-1} g_{\nu^i} \mathbf{H}_{\nu^i}(z_{\nu^k}) \ . \quad (5.2.27)$$

In the sum that appears on the right side of (5.2.27), only the terms such that $\nu^i \leq \nu^k$ are non-zero. When evaluating the computational cost in the above operation, one should make the distinction betwen the cost of the evaluation of $g(z_{\nu^k})$ and of computing the linear combination $\sum_{i=1}^{k-1} g_{\nu^i} \mathbf{H}_{\nu^i}(z_{\nu^k})$. In the case where the evaluation of g requires running a heavy numerical code (for example when $g(y)$ is an output associated with $u(y)$ the solution of a parametric PDE), the first cost dominates the second one. Once the evaluation of $g(z_{\nu^k})$ is done, the cost of the computation of g_{ν^k} amounts, upon assuming that the values $\mathbf{H}_{\nu^i}(z_{\nu^k})$ are tabulated, to the execution of $2\#(\{i : \nu^i < \nu^k\})$ usual sum and product operations. This shows that computing $\mathcal{I}_{\Lambda_n}$ by the recursive procedure is equal in cost to the ineluctable n evaluations of g at interpolation points of the grid Γ_{Λ_n} plus at most a number $K(\Lambda_n)$ of usual operations where for Λ lower, we have introducted

$$K(\Lambda) := \sum_{\nu \in \Lambda} 2\#(\mathcal{B}_\nu) \leq 2(\#(\Lambda))^2. \quad (5.2.28)$$

where we have used that $\#(\mathcal{B}_\nu) \leq \#(\Lambda)$ which results from the lower structure of Λ. If the index sets Λ_n in the context of parametric PDEs are known in advance, then the complexity of the construction of the operators $\mathcal{I}_{\Lambda_n}$ is dominated by the n evaluations of the targeted function.

The algorithm in the above lemma is also efficient to construct the operator $\mathcal{I}_\Lambda g$ for any given lower set Λ. Indeed, by iteratively removing its maximal elements, we see that any such set can be written as $\Lambda = \Lambda_k$ with $k := \#(\Lambda)$ and $(\Lambda_n)_{1 \leq n \leq k}$ a sequence of the type given in the lemma.

We have seen that the coefficients g_ν only depend on g and on the index ν and are independent on the index set Λ. These coefficients can be viewed as the unique coordinates of g in the hierarchical basis $(\mathbf{H}_\nu)_{\nu \in \mathcal{F}}$. One should however be cautious when writing the expansion

$$g = \sum_{\nu \in \mathcal{F}} g_\nu \mathbf{H}_\nu, \quad (5.2.29)$$

since it may fail to converge for certain functions g regardless of the ordering of the summation. However, it will be proved to converge for functions that can be approximated sufficiently well by polynomials, based on the stability analysis of the interpolation operator which is the object of §5.3.

5.3 The Lebesgue constant

In the construction of the previous section, any sequence $(z_k)_{k\geq 0}$ of pairwise distinct points in $[-1,1]$ can be used for the contraction of interpolation grids. However, the choice of the sequence is critical for the stability of the resulting interpolation operators $\mathcal{I}_\Lambda$, expressed by the Lebesgue constant

$$\mathbb{L}_\Lambda := \sup_{g\in B(U)\setminus\{0\}} \frac{\|\mathcal{I}_\Lambda g\|_{L^\infty(U)}}{\|g\|_{L^\infty(U)}}\,, \tag{5.3.1}$$

where $B(U)$ is the set of bounded functions g on U which are defined everywhere on U. In the case where Λ is supported in one direction, for example $\Lambda := \{0_\mathcal{F}, e_1, 2e_1, \ldots, ke_1\}$, then $\mathcal{I}_\Lambda g = I_k g_1$ with g_1 is defined over $[-1,1]$ by $g_1(t) = g(t, z_0, z_0, \ldots)$. In this case, studying the stability of $\mathcal{I}_\Lambda$ amounts to the study of the stability of I_k. We are interested in choosing sequences $(z_k)_{k\geq 0}$ such that the Lebesgue constants

$$\mathbb{L}_k = \max_{g\in C([-1,1])\setminus\{0\}} \frac{\|I_k g\|_{L^\infty([-1,1])}}{\|g\|_{L^\infty([-1,1])}}, \tag{5.3.2}$$

associated with the sections $\{z_0, \ldots, z_k\}$ have moderate growth with k. A classical example of such univariate sequences are Leja sequence (5.4.18) which numerically show moderate growth of $\mathbb{L}_k$. In addition, the choice of a Leja sequence for $(z_k)_{k\geq 0}$ has an interesting implication on the adaptive choice of the sets Λ_n as we explain in the next section.

In this section, we analyzes the Lebesgue constant of the operators $\mathcal{I}_\Lambda$. We provide bounds for theses constants and show that the stability of the operators $\mathcal{I}_\Lambda$ is indeed strongly tied to the stability of the univariate operator I_k. A crude, yet useful, way to estimate $\mathbb{L}_\Lambda$ is by using triangle inequality which gives

$$\mathbb{L}_\Lambda \leq \sum_{\nu\in\Lambda} \delta_\nu, \tag{5.3.3}$$

where we have defined for $\nu \in \mathcal{F}$

$$\delta_\nu := \sup_{g\in B(U)\setminus\{0\}} \frac{\|\Delta_\nu g\|_{L^\infty(U)}}{\|g\|_{L^\infty(U)}}, \tag{5.3.4}$$

where $B(U)$ is defined as for (5.3.1). It is readily seen that

$$\delta_\nu := \prod_{j\geq 1} \delta_{\nu_j}\,, \tag{5.3.5}$$

where

$$\delta_k := \sup_{g\in C([-1,1])\setminus\{0\}} \frac{\|\Delta_k g\|_{L^\infty([-1,1])}}{\|g\|_{L^\infty([-1,1])}} \leq \mathbb{L}_{k-1} + \mathbb{L}_k\,, \tag{5.3.6}$$

with $\mathbb{L}_k$ as in (5.3.2) and the convention that $\mathbb{L}_{-1} := 0$ since $I_{-1} := 0$. Let us remark that the product in (5.3.5) is actually finite. Indeed, since I_0 is defined by $I_0 g = g(z_0)\mathbf{1}$, then $\delta_0 = \mathbb{L}_0 = 1$. We have then

$$\mathbb{L}_\Lambda \leq \sum_{\nu\in\Lambda} \prod_{j\geq 1} (\mathbb{L}_{\nu_j - 1} + \mathbb{L}_{\nu_j}) \tag{5.3.7}$$

The bound (5.3.7) is of course crude, since we did not take advantage of the telescoping nature in the summation of the Δ_ν. For instance, when Λ is a rectangular block, i.e. $\Lambda = \mathcal{B}_\nu$ for some $\nu \in \mathcal{F}$, then we have seen that $\mathcal{I}_\Lambda = I_\nu = \otimes_{j\geq 1} I_{\nu_j}$ so that in such case, the exact value of the Lebesgue constant is giving by the smaller value

$$\mathbb{L}_{\mathcal{B}_\nu} = \prod_{j\geq 1} \mathbb{L}_{\nu_j} \ . \tag{5.3.8}$$

Nevertheless, for general lower sets Λ, we can use the bounds (5.3.3) and (5.3.7) to study the behaviour of the Lebesgue constant $\mathbb{L}_\Lambda$ as the dimension $\#(\Lambda)$ of the polynomial space $\mathbb{P}_\Lambda$ grows. The following result shows that when certain algebraic bounds are available for the $\mathbb{L}_k$ in term of $(k+1)$ the dimension of the polynomial space $\mathbb{P}_k$, then similar algebraic bounds can be derived for $\mathbb{L}_\Lambda$ in terms of $\#(\Lambda)$ regardless of the dimension d and of the shape of Λ.

Lemma 5.3.1

If the Lebesgue constants $\mathbb{L}_k$ satisfy

$$\mathbb{L}_k \leq (k+1)^\theta, \quad k \geq 0, \tag{5.3.9}$$

for some $\theta \geq 1$, then for any monotone set Λ, one has

$$\mathbb{L}_\Lambda \leq (\#\Lambda)^{\theta+1} \tag{5.3.10}$$

Proof: If $\theta \geq 1$, then for any $k \geq 0$ one has $\mathbb{L}_k + \mathbb{L}_{k-1} \leq (k+1)^\theta + k^\theta \leq (2k+1)(k+1)^{\theta-1}$, therefore, for $\nu \in \Lambda$

$$\begin{aligned} \prod_{j\geq 1}(\mathbb{L}_{\nu_j} + \mathbb{L}_{\nu_j - 1}) &\leq \Big(\prod_{j\geq 1}(\nu_j + 1)\Big)^{\theta-1} \prod_{j\geq 1}(2\nu_j + 1) \\ &= (\#(\mathcal{B}_\nu))^{\theta-1} \prod_{j\geq 1}(2\nu_j + 1) \\ &\leq (\#(\Lambda))^{\theta-1} \prod_{j\geq 1}(2\nu_j + 1), \end{aligned}$$

where we have used $\mathcal{B}_\nu \subset \Lambda$ which follows from Λ being a lower set. To complete the proof, it remains to show that $K_{0,0}(\Lambda) \leq (\#\Lambda)^2$, where

$$K_{0,0}(\Lambda) := \sum_{\nu\in\Lambda} \prod_{j\geq 1} (2\nu_j + 1) \ .$$

This is proved is the appendix, Lemma A.4.1. ■

Remark 5.3.2

In order to establish the result in Lemma 5.3.1, one only needs to have

$$\delta_k \leq (2k+1)(k+1)^{\theta-1}. \tag{5.3.11}$$

In the case where $(z_k)_{k\geq 0}$ is a Leja sequence defined by (5.4.18) for some initial point $z_0 \in [-1,1]$, the hierarchical polynomials h_k defined by (5.2.19) satisfy $|h_k(t)| \leq |h_k(z_k)| = 1$ for any $t \in [-1,1]$. Since, according to (5.2.18), we have

$$\Delta_k g = (g(z_k) - I_{k-1}g(z_k))h_k. \tag{5.3.12}$$

It follows that

$$\delta_k \leq 1 + \mathbb{L}_{k-1}, \tag{5.3.13}$$

and

$$\mathbb{L}_\Lambda \leq \sum_{\nu\in\Lambda}\prod_{j\geq 1}(1+\mathbb{L}_{\nu_j-1}) \tag{5.3.14}$$

which are improvements over (5.3.6) and (5.3.14) and can be used to prove (5.3.1) using only that

$$1 + \mathbb{L}_{k-1} \leq (2k+1)(k+1)^{\theta-1}. \tag{5.3.15}$$

Let us observe that since $\mathbb{L}_{-1} = 0$ and $\mathbb{L}_0 = 1$, bounds of the form $\mathbb{L}_k \leq (k+1)^\theta$ can be established for some $\theta \geq 1$ provided that $\mathbb{L}_k$ are bounded as $\mathcal{O}((k+1)^\theta)$ for some $\theta > 0$. Such bounds are available for sequences that we discuss in the next chapter.

We don't have a simple argument for adapting the result of Lemma 5.3.1 when the Lebesgue constant $\mathbb{L}_k$ grows logarithmically, i.e. $\mathbb{L}_k \lesssim \log(k+1)$ for $k \geq 1$. We are however able, using similar arguments as above, to bound Lebssgue constant for sparse grids based on Clenshaw-Curtis points.

In the case of sparse grids interpolation operator as defined in (5.2.10), we can take benefit from the telescopic sums in order to sharpen bound for the Lebesgue constant $\mathbb{L}_{m(\Lambda)}$ of $\mathcal{I}_{m(\Lambda)}$ as in (5.2.10). Indeed, one has

$$\mathbb{L}_{m(\Lambda)} \leq \sum_{i\in\Lambda}\prod_{j\geq 1}(\mathbb{L}_{m(i_j)} + \mathbb{L}_{m(i_j-1)}), \tag{5.3.16}$$

In particular if m is a doubling rule, that is $m(-1) = -1$, $m(0) = 0$ and $m(i) = 2^i$ for any $i \geq 1$, and the point of interpolation used for $I_{m(i)}$ are the Clenshew-Curtis abscissas of order 2^i, i.e.

$$\cos\Big(\frac{j\pi}{2^i}\Big), \qquad j = 0,\dots,2^i, \tag{5.3.17}$$

then since

$$\mathbb{L}_{m(i)} \leq \frac{2}{\pi}\log(2^i) + 1 = \frac{2\log 2}{\pi}i + 1 \leq i + 1 \tag{5.3.18}$$

we deduce that

$$\mathbb{L}_{m(\Lambda)} = \sum_{i\in\Lambda}\prod_{j\geq 1}(2i_j+1) \leq (\#(\Lambda))^2. \tag{5.3.19}$$

The Lebesgue constant is moderate in view of the dimension of $\mathbb{P}_{m(\Lambda)}$ which is equal to

$$\#(m(\Lambda)) = \sum_{i\in\Lambda} 2^{|i|}. \tag{5.3.20}$$

5.4 Application of high dimensional interpolation to parametric PDEs

5.4.1 Interpolation of Banach valued functions

We are interested in applying our interpolation process to the solution map $y \mapsto u(y)$ defined by exact or approximate resolution of the parametric PDE (5.1.1) for the given parameter y. Therefore, we want to interpolate a function which is not real or complex valued, but instead takes values in the solution space V. The generalization of the interpolation operator $\mathcal{I}_\Lambda$ to this setting is straighforward: $\mathcal{I}_\Lambda u$ is the unique function in $\mathbb{V}_\Lambda$ that coincides with u at the points $\{z_\nu\}_{\nu\in\Lambda}$. As in the scalar case, it can be expanded according to

$$\mathcal{I}_\Lambda u = \sum_{\nu\in\Lambda} u_\nu \mathbf{H}_\nu \ , \tag{5.4.1}$$

where the coefficients $u_\nu \in V$ can be computed in a recursive way similar to (5.2.27):

$$u_{\nu^1} = u(z_{0_\mathcal{F}}), \quad u_{\nu^k} = u(z_{\nu^k}) - \sum_{i=1}^{k-1} u_{\nu^i}\mathbf{H}_{\nu^i}(z_{\nu^k}), \tag{5.4.2}$$

where $\Lambda_n = \{\nu^1,\dots,\nu^n\}$, $n = 1,2,\dots$, is a nested sequence of lower sets. We are interested in the accuracy of the interpolant in the sense of the maximum error

$$\|u - \mathcal{I}_\Lambda u\|_{\mathcal{V}_\infty} := \sup_{y\in U}\|u(y) - \mathcal{I}_\Lambda u(y)\|_V. \tag{5.4.3}$$

The same reasoning as for interpolation of scalar valued functions shows that

$$\|u - \mathcal{I}_\Lambda u\|_{\mathcal{V}_\infty} \leq (1+\mathbb{L}_\Lambda)\inf_{v\in\mathbb{V}_\Lambda}\|u-v\|_{\mathcal{V}_\infty}, \tag{5.4.4}$$

where $\mathbb{L}_\Lambda$ is the Lebesgue constant associated to the interpolation operator $\mathcal{I}_\Lambda$ which was defined and studied in the previous section.

5.4.2 Convergence rates for a parametric, elliptic model problem

As already explained in the introduction, for the model elliptic problem (1.1.1), one can establish convergence rates in $\mathcal{V}_\infty$ and $\mathcal{V}_2$ where $V = H_0^1(D)$, for polynomial approximation that are robust with respect to the parametric dimension. This is also the case for a large class of parametric models as we have shown in Chapter 2 with other choices of the space V. Since the model in Chapter 1 can be viewed as a particular case of the frameworks in Chapter 2, we only work with the notation of the latter involving only Legendre polynomial.

Without going into details, the result of Chapter 2 is stated as following: If the parametric PDE of the general form (5.1.1) is well posed in some Banach space V for any $y \in U$ and the operator $\mathcal{D}$ satisfies a (p, ε)-holomorphy assumption, see Definition 2.2.1, then the sequence $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ of Legendre coefficients associated with the the Legendre family $(P_\nu)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$. Hence as in (2.2.7), there exists a sequence $(\Lambda_n)_{n\geq 1}$ of nested lower sets such that $\#(\Lambda_n) = n$ and

$$\inf_{v\in\mathbb{V}_{\Lambda_n}} \|u - v\|_{\mathcal{V}_\infty} \leq C(n+1)^{-s}, \quad s := \frac{1}{p} - 1. \tag{5.4.5}$$

where $C > 0$ does not depend on n. One way to compute an approximation of the solution map u in the polynomial spaces $\mathbb{V}_{\Lambda_n}$ is by using the incremental interpolation process we introduced in the previous sections. One way to study the rate of convergence of $\mathcal{I}_{\Lambda_n} u$ towards u is through the analysis of stability using Lebesgue constant. Combining (5.4.4) and (5.4.5), we obtain

$$\|u - \mathcal{I}_{\Lambda_n} u\|_{\mathcal{V}_\infty} \leq C(1 + \mathbb{L}_{\Lambda_n})(n+1)^{-s}\,. \tag{5.4.6}$$

We have seen in §5.3 that the Lebesgue constant can be controlled by a bound of the form

$$\mathbb{L}_{\Lambda_n} \leq (\#(\Lambda_n))^{\theta+1} = n^{\theta+1}, \tag{5.4.7}$$

when the univariate sequence $(z_k)_{k\geq 0}$ is chosen so that the growth of the Lebesgue constants associated with the sections $\{z_0, \ldots, z_k\}$ satisfy $\mathbb{L}_k \leq (k+1)^\theta$ for some $\theta \geq 1$. Sequences with the previous property are studied in the next chapter for which the value θ is proved to be smaller than 2. Using such sequences, we thus obtain a convergence estimate of the form

$$\|u - \mathcal{I}_{\Lambda_n} u\|_{\mathcal{V}_\infty} \leq C(n+1)^{\theta+1-s}. \tag{5.4.8}$$

With this simple stability (via the bound for the Lebesgue constant) plus consistency (via the n-term approximation result) analysis, the convergence rate obtained in (5.4.8) is deteriorated at worse by $(\theta + 1)$ compared to the n-term approximation rate s in (5.4.5). Using only this analysis, one can not say if a convergence is guaranteed when $s \leq \theta + 1$.

The following result recovers the best n-term approximation rate $\mathcal{O}((n+1)^{-s})$ for the interpolation based on a different choice of lower sets than the sequence (Λ_n) above. This analysis is similar to an analysis that was developed in [70, 69, 7, 26] in the particular case of the solution map u of elliptic model (1.1.1) with affine dependence and uniform ellipticity assumption. It is based on the fact that the algebraic growth of the univariate Lebesgue constants $\mathbb{L}_k$ can be absorbed inside the estimates obtained for Legendre or Taylor coefficients based on analyticity.

Lemma 5.4.1

Assume that $u = \sum_{\nu\in\mathcal{F}} u_\nu P_\nu$ in the sense of unconditional convergence in $\mathcal{V}_\infty$. If the univariate sequence $(z_k)_{k\geq 0}$ is chosen so that $\mathbb{L}_k \leq (k+1)^\theta$, for some $\theta \geq 0$, then for any lower set Λ, one has

$$\|u - \mathcal{I}_\Lambda u\|_{\mathcal{V}_\infty} \leq 2 \sum_{\nu\notin\Lambda} p_\nu(\theta) \|u_\nu\|_V \ , \tag{5.4.9}$$

where

$$p_\nu(\theta) := \prod_{j\geq 1} (1+\nu_j)^{\theta+1} \ . \tag{5.4.10}$$

Proof: The unconditional convergence in $\mathcal{V}_\infty$ of the Legendre series yields that for any lower set Λ,

$$\mathcal{I}_\Lambda u = \mathcal{I}_\Lambda\Big(\sum_{\nu\in\mathcal{F}} u_\nu P_\nu\Big) = \sum_{\nu\in\mathcal{F}} u_\nu \mathcal{I}_\Lambda P_\nu = \sum_{\nu\in\Lambda} u_\nu \mathcal{I}_\Lambda P_\nu + \sum_{\nu\notin\Lambda} u_\nu \mathcal{I}_\Lambda P_\nu \ .$$

The univariate polynomial P_k is of degree k, therefore for any $\nu \in \mathcal{F}$, the polynomial P_ν belongs to $\mathbb{P}_{\mathcal{B}_\nu}$ where $\mathcal{B}_\nu := \{\mu \in \mathcal{F} : \mu \leq \nu\}$. If $\nu \in \mathcal{F}$, the lower structure of Λ implies that $\mathcal{B}_\nu \subset \Lambda$, hence $P_\nu \in \mathbb{P}_\Lambda$, so that $\mathcal{I}_\Lambda P_\nu = P_\nu$. For $\nu \notin \Lambda$, since we have $\Delta_\mu P_\nu = \prod_{j\geq 1} \Delta_{\mu_j} P_{\nu_j} = 0$ for any $\mu \not\leq \nu$, then for any $\nu \in \mathcal{F}$ one has

$$\mathcal{I}_\Lambda P_\nu = \sum_{\mu\in\Lambda:\mu\leq\nu} \Delta_\mu P_\nu = \mathcal{I}_{\Lambda\cap\mathcal{B}_\nu} P_\nu \ .$$

The two previous observations imply

$$u - \mathcal{I}_\Lambda u = \sum_{\nu\notin\Lambda} u_\nu (\mathcal{I} - \mathcal{I}_{\Lambda\cap\mathcal{B}_\nu}) P_\nu,$$

where $\mathcal{I}$ denotes the identity operator defined over $C(U)$. Therefore

$$\|u - \mathcal{I}_\Lambda u\|_{\mathcal{V}_\infty} \leq \sum_{\nu\notin\Lambda} \|u_\nu\|_V (1 + \mathbb{L}_{\Lambda\cap\mathcal{B}_\nu}) \|P_\nu\|_{L^\infty(U)} \leq 2 \sum_{\nu\notin\Lambda} \|u_\nu\|_V \mathbb{L}_{\Lambda\cap\mathcal{B}_\nu} \ .$$

where we have used that the Legendte polynomials have infinite norms equal to 1 and the Lebesgue constant always greater than 1. If the univariate sequence is such that

$\lambda_k \leq (k+1)^\theta$ for some $\theta > 0$, then we have

$$\mathbb{L}_{\Lambda\cap\mathcal{B}_\nu} \leq \#(\Lambda\cap\mathcal{B}_\nu)^{\theta+1} \leq \#(\mathcal{B}_\nu)^{\theta+1} = \Big(\prod_{j\geq 1}(1+\nu_j)\Big)^{\theta+1} = p_\nu(\theta),$$

so that

$$\|u - \mathcal{I}_\Lambda u\|_{\mathcal{V}_\infty} \leq 2\sum_{\nu\notin\Lambda} p_\nu(\theta)\|u_\nu\|_V\ ,$$

which completes the proof. ■

The above lemma can be generalized to any polynomial expansion other than Legendre series, for example the expansion into Taylor polynomials, provided that unconditional convergence holds and $(P_\nu)_{\nu\in\mathcal{F}}$ any family of tensorized polynomials, i.e. $P_\nu = \otimes_{j\geq 1} P_{\nu_j}$, such that $(P_k)_{k\geq 0}$ is a family of univariate polynomials with $P_0 = 1$, P_k is of degree k and $\|P_k\|_{L^\infty([-1,1])} = 1$. This is in particular the case with the elliptic linear model studied in Chapter 1. We focus here on the Legendre series, which allows us to use the results of Chapter 2, applicable to more general parametric PDEs. We have seen that unconditional convergence holds under assumptions on the anisotropic dependence of the parametric PDE (5.1.1) on y and it is based on explicit bounds for the V-norms of Legendre coefficients. These bounds are obtained by application of the Cauchy formula, on the holomorphy extension of the solution map u as in the proof of Theorem 1.6.9 and are of the form (2.2.5)

$$\|v_\nu\|_V \leq \Big(\prod_{j\geq 1}(\nu_j+1)\Big)\, C_\varepsilon \inf_{\rho\in\mathcal{A}_{\varepsilon,b}} \Big\{\rho^{-\nu}\prod_{j\geq 1:\nu_j\neq 0}\varphi(\rho_j)\Big\}, \qquad \nu\in\mathcal{F} \tag{5.4.11}$$

where C_ε is a constant not depending on ν and $\mathcal{A}_{\varepsilon,b}$ denote the set of sequences $\rho = (\rho_j)_{j\geq 1}$ which are (b,ε)-admissible, see (2.2.2). Given now the solution map u which is a sum of its Legendre series with algebraic rates and $\theta \geq 1$, we introduce the sequence $(\alpha_\nu)_{\nu\in\mathcal{F}}$ defined by

$$\alpha_\nu = p_\nu(\theta)\|u_\nu\|_V, \qquad \nu\in\mathcal{F} \tag{5.4.12}$$

where $p_\nu(\theta)$ is as in the previous lemma. In view of the relation between Legendre coefficients $\|v_\nu\|_V$ and $\|u_\nu\|_V$, the sequence α is bounded according to $\alpha_\nu \leq C_\varepsilon C_\nu(\theta) g_\nu$ for any $\nu\in\Lambda$, where $(g_\nu)_{\nu\in\mathcal{F}}$ defined as in Chapter 1, formula (1.6.15) and

$$C_\nu(\theta) := \Big(\prod_{j\geq 1:\nu_j\neq 0}\sqrt{2}(\nu_j+1)^{\theta+\frac{5}{2}}\Big), \tag{5.4.13}$$

which is of the form of what we have called a multi-dimensional algebraic deterioration in Chapter 1, formula (1.3.26). In view of Theorem 1.6.5, we have then the following,

Theorem 5.4.2

Assume that the differential operator $\mathcal{D}$ satisfies the (p,ε) holomorphy assumption

for some $0 < p < 1$ and $\varepsilon > 0$. Then for any $\theta \geq 1$ and for $p_\nu(\theta)$ as in (5.4.10), the sequence $(p_\nu(\theta)\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ belongs to $\ell^p_m(\mathcal{F})$.

We have thus established the following convergence result.

Theorem 5.4.3

Assume that $\mathcal{D}$ satisfies the (p,ε)-holomorphy for some $0 < p < 1$ and $\varepsilon > 0$. If the univariate sequence $(z_k)_{k\geq 0}$ is chosen so that the growth of the Lebesgue constant satisfies $\mathbb{L}_k \leq (k+1)^\theta$ for some $\theta > 0$, then there exists a sequence $(\Lambda_n)_{n\geq 1}$ of lower sets Λ_n such that $\#(\Lambda_n) = n$ and

$$\|u - \mathcal{I}_{\Lambda_n} u\|_{\mathcal{V}_\infty} \leq C(n+1)^{-s}, \quad s = \frac{1}{p} - 1\,. \tag{5.4.14}$$

5.4.3 Adaptive selection of polynomial spaces

We now discuss the adaptive selection of a nested sequence $(\Lambda_n)_{n\geq 1}$. Let us begin with the following analogy: if $(\mathbf{H}_\nu)_{\nu\in\mathcal{F}}$ was an orthonormal basis of $L^2(U)$ then the choice of an index set Λ_n that minimize the L^2 error when truncating the expansion (5.2.29) would be the indices corresponding to the n largest $|g_\nu|$.

In our current setting however, $(\mathbf{H}_\nu)_{\nu\in\mathcal{F}}$ is not an orthonormal basis and we are rather interested in controlling the error in an infinite sense. A possible greedy strategy is to define Λ_n as the set of indices corresponding to the n largest contributions of (5.2.29) measured in the L^∞ metric, i.e. the n largest $a_\nu|g_\nu|$, where

$$a_\nu := \|\mathbf{H}_\nu\|_{L^\infty(U)} := \prod_{j\geq 1} \|h_{\nu_j}\|_{L^\infty([-1,1])}. \tag{5.4.15}$$

This strategy obviously can give rise to a nested sequence $(\Lambda_n)_{n\geq 1}$, however the sets Λ_n are not ensured to be lower. In addition, it is not computationally feasible since finding the n largest contributions in (5.2.29) hints that we should have computed all contributions. In order to correct these defects, we define for any lower set Λ a set of neighbours

$$\mathcal{N}(\Lambda) := \Big\{\nu \notin \Lambda \;:\; \Lambda \cup \{\nu\} \;\text{ lower}\Big\}. \tag{5.4.16}$$

This set consists of those $\nu \notin \Lambda$ satisfying $\nu - e_j \in \Lambda$ for any j such that $\nu_j \neq 0$. We remark that a natural variant of the first strategy, that leads to a nested sequence of lower sets, is the following greedy adaptive algorithm that we call *adaptive interpolation (AI) algorithm.*

Algorithm 5.4.4

- *Start with* $\Lambda_1 := \{0_{\mathcal{F}}\}$.
- *Assuming that* Λ_{n-1} *has been computed, build* $\mathcal{N}(\Lambda_{n-1})$*, compute* g_ν *for any* $\nu \in \mathcal{N}(\Lambda_{n-1})$ *and find*

$$\nu^n := \operatorname{argmax}\Big\{ a_\nu |g_\nu| \ : \ \nu \in \mathcal{N}(\Lambda_{n-1}) \Big\}, \tag{5.4.17}$$

- *Set* $\Lambda_n = \Lambda_{n-1} \cup \{\nu^n\}$.

Let us observe that since $\mathbf{H}_\nu(z_\nu) = 1$, we obviously have that $a_\nu \geq 1$. On the other hand, when $(z_k)_{k\geq 0}$ is a *Leja sequence* on $[-1,1]$ built according to $z_0 \in [-1,1]$ and the inductive construction

$$z_k := \operatorname{Argmax}_{z\in P} \prod_{j=0}^{k-1} |z - z_j|, \tag{5.4.18}$$

we obviously have $\max_{z\in[-1,1]} |h_k(z)| = |h_k(z_k)| = 1$ and therefore

$$a_\nu = \mathbf{H}_\nu(z_\nu) = 1. \tag{5.4.19}$$

In such a case, in view of (5.2.27), the greedy strategy (5.4.17) amounts in choosing the new index in $\mathcal{N}(\Lambda_{n-1})$ that maximizes the interpolation error at the corresponding new grid point:

$$\nu^n := \operatorname{argmax}\Big\{ |g(z_\nu) - \mathcal{I}_{\Lambda_{n-1}} g(z_\nu)| \ : \ \nu \in \mathcal{N}(\Lambda_{n-1}) \Big\}. \tag{5.4.20}$$

Similarly to the algorithm "Largest Neighbour" described in Chapter 3, the algorithm AI has many computational advantages. The quantities a_ν and g_ν depends only on ν and are then computed for every multi-index once and for all. Although $\mathcal{N}(\Lambda_n)$ is of infinite cardinality when $d = \infty$, the indices that update $\mathcal{N}(\Lambda_{n-1})$ to $\mathcal{N}(\Lambda_n)$ are of finite number smaller that $\#(\operatorname{supp}\Lambda_{n-1})$. One then only needs to compute a_ν and g_ν for the infinite set

$$\mathcal{N}(\{0_{\mathcal{F}}\}) := \Big\{ e_j : j \geq 1 \Big\} \tag{5.4.21}$$

which are giving by

$$a_{e_j} = \frac{1+|z_0|}{|z_1 - z_0|}, \qquad g_{e_j} = |g(z_{e_j}) - g(z_{0_{\mathcal{F}}})| = |g(z_0,\ldots,z_0,z_1,z_0,\ldots) - g(z_0,z_0,\ldots)|. \tag{5.4.22}$$

One can then either have access to all e_j or use an argument such as mean value theorem in order to deduce a-priori estimates on the quantities $|g_{e_j}|$ that can be used as hints in order to make the algorithm feasible. We provide bounds for $|g_{e_j}|$ in the next section for the particular setting of elliptic parametric PDEs.

The greedy strategy has also several defects. The first one is that it may simply fail to converge, even if there exist sequences $(\Lambda_n)_{n\geq 0}$ such that $\mathcal{I}_{\Lambda_n} g$ converges to g at a high rate. This is due to data oscillation that could return an artificially small interpolation error at the new grid point. Consider for example a two dimensional function of the form

$$g(y) = g(y_1, y_2) = g_1(y_1)g_2(y_2), \tag{5.4.23}$$

where g_1 and g_2 are non-polynomial smooth functions such that g_2 takes the same values at the points z_0 and z_1. Then, the algorithm will select sets Λ_n that consist of the indices $\nu = (k, 0)$ for $k = 0, \ldots, n-1$, since the interpolation error at the point $z_{(k,1)} = (z_k, z_1)$ will always be null. Although this type of situation might be viewed as pathological, it reflects the fact that the algorithm might fail in its first steps to identify the significant variables. One way to avoid this is to impose that when all interpolation errors $|g(z_\nu) - \mathcal{I}_{\Lambda_{n-1}} g(z_\nu)|$ for $\nu \in \mathcal{N}(\Lambda_n)$ are smaller than some prescribed tolerance $\varepsilon_n > 0$ (that is either fixed or tends to 0 as n grows), then the new index ν^n is chosen arbitrarily from $\mathcal{N}(\Lambda_n)$.

The second defect which we already encountered is that in the infinite dimensional framework $d = \infty$, the set of neighbours $\mathcal{N}(\Lambda)$ has infinite cardinality. One way to treat this defect is by modifying the algorithm and choosing at each iteration k, the next element in the reduced set of neighbors

$$\mathcal{N}_J(\Lambda) := \Big\{ \nu \in \mathcal{N}(\Lambda) : \nu_j = 0 \ \text{ if } \ j > J + 1 \Big\}, \tag{5.4.24}$$

where $J = J(\Lambda)$ is such that $\nu_j = 0$ for any $\nu \in \Lambda$ and $j \geq J$. This means that we can activate at most one new variable at each iteration step.

Even with such modifications, it is not clear to understand under which additional assumptions on g the adaptive greedy selection procedure picks sets $(\Lambda_n)_{n\geq 0}$ such that the interpolation $\mathcal{I}_{\Lambda_n} g$ has a guaranteed convergence rate comparable to that of an optimal choice of sets. We give in §5 several numerical examples that illustrate the good practical behaviour of this algorithm.

5.5 Numerical experiments

5.5.1 Scalar valued functions

We first consider the interpolation of scalar valued functions $u : U \to \mathbb{R}$ where now $U = [-1, 1]^d$. Our objective is to test the adaptive algorithm AI proposed in §5.4.3 in various ways:

- Ability to select good lower index sets Λ_n, in particular when the function has anisotropic dependence on the variables.

- The effect of the choice of the univariate sequence $(z_k)_{k\geq 0}$, in particular on the robustness of the interpolation with respect to noise in the measurements.
- Robustness of the performance with respect to the dimension d when the function depends only on few unknown variables, or when the dependence with respect to the variables is sufficiently anisotropic.

We consider three possible choices for the univariate sequence $(z_k)_{k\geq 0}$:

- **Uniform sequence** (Q): $z_0 = 1$, $z_1 = -1$, $z_2 = 0$ and for $k > 1$ we set $z_{2k+1} = \frac{1}{2}\sum_{j=0}^{n} \varepsilon_j 2^{-j}$ where $k = \sum_{j=0}^{n} \varepsilon_j 2^j$ is the binary expansion of k and $z_{2k+2} = -z_{2k+1}$. Such a choice produces a uniform subdivision of $[-1, 1]$ of step size 2^{-j} for the particular sections $(z_0, \ldots, z_{2^j})$, and avoids accumulation of points on a region of the interval for the intermediate sections $(z_0, \ldots, z_k)$, $2^j < k < 2^{j+1}$.
- **Leja sequence** (L): $z_0 = 1$ and the sequence z_k is defined recursively on $[-1, 1]$ by (5.4.18). Here also we have $z_1 = -1$ and $z_2 = 0$.
- **$\Re$-Leja sequence** (R): this is the projection on $[-1, 1]$ of a Leja sequence for the complex unit disk initiated at 1. The $\Re$-Leja sequence has an explicit structure which is very similar to that of the sequence Q in the sense that $(z_0, \ldots, z_{2^j})$ are Clenshaw-Curtis abscissas, that is the projections on the real axis of a uniform subdivision of the upper half-circle with end-points at -1 and 1. This is explained in details in the next chapter. The $\Re$-Leja sequence we use here has an explicit formula. Namely $z_0 = 1$, $z_1 = -1$ and for $2^n \leq k < 2^{n+1}$ having the binary expansion $k = 2^n + \sum_{j=0}^{n-1} a_j 2^j$,

$$z_{k+1} = \cos\Big(\frac{\pi}{2^{n+1}} + \pi \sum_{j=0}^{n-1} a_j 2^{-j}\Big). \tag{5.5.1}$$

Our first example is the function of $d = 16$ variables

$$u_1(y) = u(y_1, \ldots, y_{16}) = y_3 \sin(y_4 + y_{16}), \tag{5.5.2}$$

that in fact depends only on 3 variables. Figure 5.5.1 displays the uniform error between u and $\mathcal{I}_{\Lambda_n} u$ in terms of $n = \#(\Lambda_n)$ for the AI algorithm based on 3 the possible choices Q, L and R for the univariate sequence $(z_k)_{k\geq 0}$.

We observe that the error decays fastly and reaches machine precision with the choices L and R for the univariate sequence, but not with the choice Q, although u_1 is analytic over $\mathbb{R}^d$. Inspection of the index sets Λ_n generated by the algorithm for the three choices reveals that for $n = 10^3$, the polynomial degree in the active variables y_4 and y_{16} reaches values above 30. This explains the bad behaviour of the error for the choice Q. Indeed, it is well known that the univariate Lebesgue constant associated to

k uniformly spaced points is higher than 2^k and therefore the amplification of machine precision measurement noise begins to deteriorate the precision. This does not occur with the choices L and R for which the Lebesgue constant has moderate growth. For these sequences, the algorithm identifies the three active variables (y_3, y_4, y_{16}) in the sense that all chosen indices ν have $\nu_j = 0$ for $j \neq 3, 4, 16$.

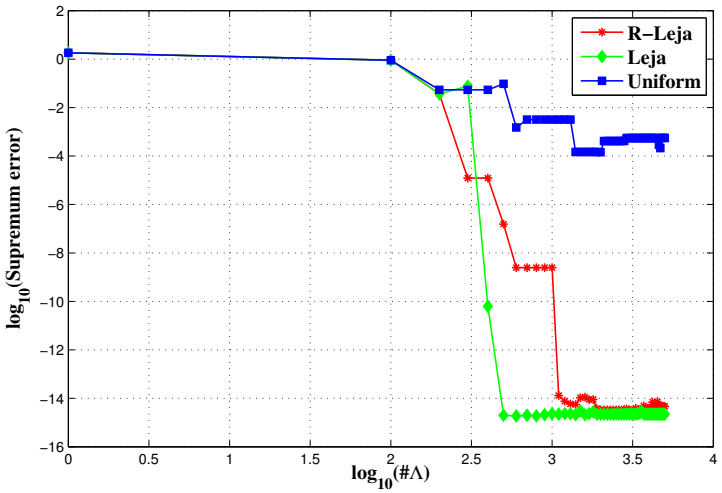

Figure 5.5.1: Uniform error of the AI algorithm applied to u_1 based on the sequences R, L and U.

In our next example, we consider the function

$$u_2(y) = \left(1 + \sum_{j=1}^{d} \gamma_j y_j\right)^{-1}, \quad \gamma_j := \frac{3}{5j^3}. \tag{5.5.3}$$

This function now depends on all variables $y_1, \ldots, y_d$ but in an strongly anisotropic way due to the decay of the weights γ_j. Since $\sum_{j=1}^{\infty} \gamma_j \approx 0.72 < 1$, the function u is analytic on U in each variable around 0, regardless of the dimension d which can be even infinite $d = \infty$. Moreover, the same analysis used in chapters 1 and 2 to prove (5.1.4) for parametric PDEs based on holomorphy arguments, shows that since $(\gamma_j)_{j\geq 1} \in \ell^p$ for any $\frac{1}{3} < p < 1$, then there exists a sequence $(\Lambda_n^*)_{n\geq 1}$ with $n = \#(\Lambda_n^*)$ such that

$$\inf_{v \in \mathbb{P}_{\Lambda_n^*}} \|u_2 - v\|_{L^\infty(U)} \leq C_{\gamma,p}(n+1)^{-s}, \quad s = \frac{1}{p} - 1, \tag{5.5.4}$$

where $C_{\gamma,p}$ is *independent* of d. Figure 5.5.2 reveals that this robustness with respect to d is also observed when using the AI algorithm (here based on the sequence R), since its convergence behaviour is almost unchanged for $d = 8$, 16, 32 and 64.

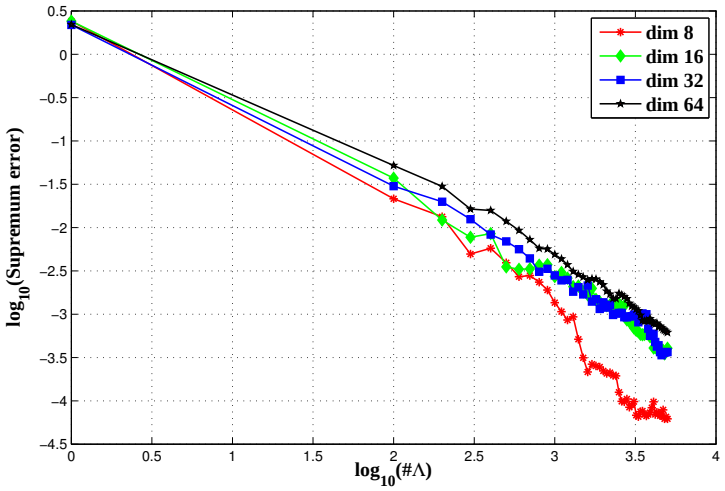

Figure 5.5.2: Uniform error for the AI algorithm applied to u_2 based on the sequence R for the dimensions $d = 8, 16, 32, 64$.

We next fix $d = 16$ and compare the error of the AI algorithm applied to u_2 based on the three sequences L, R and Q. Figure 5.5.3 reveals that, in contrast to the function u_1, the uniform sequence Q gives as good results as the sequences L and R. This can be explained by inspecting more closely the index sets Λ_n, for which one finds that for $n = 10^4$ the highest polynomial degree attained on the most active variable y_1 is 17 (due the presence of many active variables) and therefore the amplification of the machine precision noise is not yet visible.

We perform the same test with a higher additive noise, by interpolating the values $u_2(z_\nu)+\varepsilon_\nu$ where ε_ν are independent realizations of a random variable with uniform law on $[-10^{-3}, 10^{-3}]$. Figure 5.5.4 reveals that the error diverges when using the uniform sequence Q, while it decays when using R or L (however not reaching arbitrarily small values due to presence of the noise).

Finally, we consider with $d = 16$ the function

$$u_3(y) = \Big(1 + (\sum_{j=1}^{d} \gamma_j y_j)^2\Big)^{-1}, \quad \gamma_j := \frac{5}{j^3}. \tag{5.5.5}$$

Similar to u_2, this function has an anisotropic behaviour. However, in contrast to u_2, it is not analytic in each variable around 0 due to the fact that $\sum_{j=1}^{d} \gamma_j > 1$. As a result, algorithm AI based on the uniform sequence Q does not converge, even in the noiseless case, as illustrated by Figure 5.5.5. This can be viewed as a manifestation of the well known Runge phenomenon.

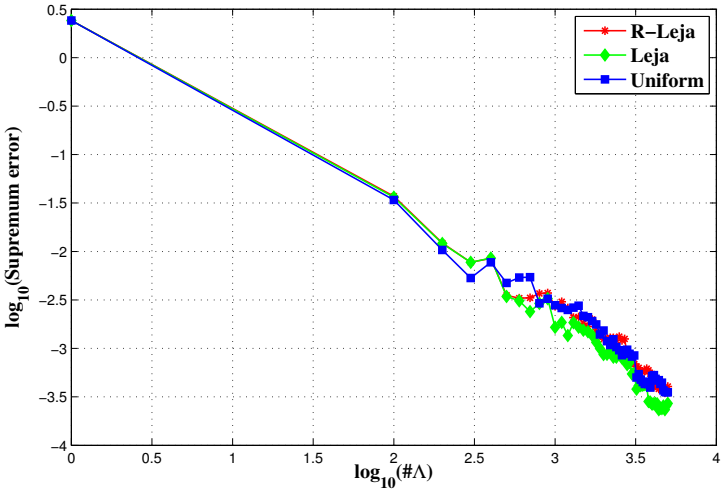

Figure 5.5.3: Uniform error of algorithm AI applied to u_2 with $d = 16$ based on the sequences R, L and Q.

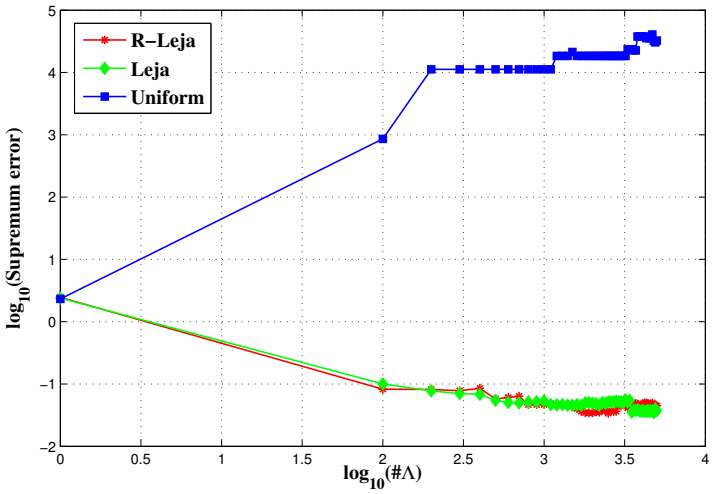

Figure 5.5.4: uniform error of algorithm AI applied to noisy evaluations of u_2 with $d = 16$ based on the sequences R, L and Q.

In summary, algorithm AI takes advantage of an anisotropic dependence on the variables, however its success is critically tied to the choice of the univariate sequence $(z_k)_{k\geq 0}$ in either one of these situations:

(i) the polynomial degree reaches high values in certain variables,

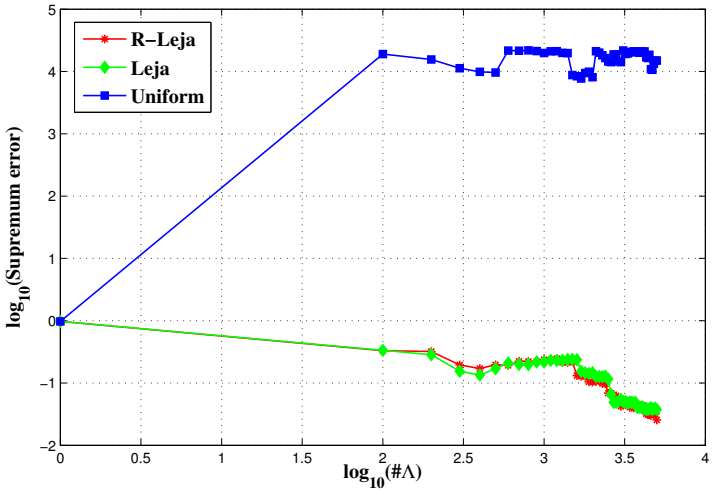

Figure 5.5.5: Uniform error of algorithm AI applied to u_3 with $d = 16$ based on the sequences R, L and Q.

(ii) the measurements are noisy,

(iii) the function has not enough smoothness in certain variables.

In all cases, both sequences R and L are good choices.

5.5.2 Parametric PDE's

We now turn to the interpolation of functions, $u : U \to V$ defined as the solution map of a parametric PDE (5.1.1) where V is the solution space. In practice, the PDE is solved by a numerical technique such as the finite element method applied with a certain mesh, and therefore we rather interpolate the numerical solution map

$$u_h : U \to V_h, \tag{5.5.6}$$

where V_h is finite dimensional.

In Chapter 3, several adaptive algorithms based on the Taylor partial sums were proposed, analyzed and implemented for the linear model elliptic PDE (1.1.1) studied in Chapter 1. The most practical and efficient of these algorithms acts in a very similar way as algorithm AI, in the sense that the set Λ_{n+1} is defined by adding to Λ_n the index ν that maximizes the V-norm of the Taylor coefficient t_ν among the set of neighbours $\mathcal{N}(\Lambda_n)$. In the sequel, we refer to this adaptive algorithm as Largest Neighbour Taylor (LNT). Note that, in contrast to AI, algorithms based on the computation of the Taylor series

such as LNT are by essence intrusive and strongly benefit from the particular structure of the problem (1.1.1): a linear equation with affine dependence of the operator on the parameters.

We compare the two algorithms AI and LNT when applied to (1.1.1) with $D = [0,1]^2$ and diffusion coefficient $a(x,y)$ given by an expansion in the two dimensional Haar wavelet basis similar to Test 2 in Chapter 3, namely

$$a(x,y) := \overline{a}(x) + \sum_{l=0}^{L} \beta_l \sum_{i=1}^{3} \sum_{k \in \{0,\dots,2^l-1\}^2} y_{l,k,i} h^i_{l,k}(x), \quad \bar{a} = 1. \tag{5.5.7}$$

In the above expansion,

$$h^i_{l,k}(x) := h^i(2^l x - k), \;\; l \in \mathbb{N}, \;\; k = (k_1,k_2) \in \{0,\dots,2^l-1\}^2, \;\; i = 1,2,3, \tag{5.5.8}$$

where the generating wavelets h^i are defined by

$$h^1(x_1,x_2) := \varphi(x_1)h(x_2), \qquad h^2(x_1,x_2) := h(x_1)\varphi(x_2) \quad \text{and} \quad h^3(x_1,x_2) := h(x_1)h(x_2), \tag{5.5.9}$$

with $\varphi := \chi_{[0,1]}$ and $h := \chi_{[0,1/2[} - \chi_{[1/2,1[}$. Using the relabelling

$$\psi_j := \beta_l h^i_{l,k} \;\text{ and }\; y_j := y_{l,k,i}, \;\text{ when }\; j = 2^{2l} + 3(2^l k_1 + k_2) + i - 1, \tag{5.5.10}$$

we may rewrite the above expansion (5.5.7) in the form $a(x,y) := \overline{a}(x) + \sum_{j=1}^{d} y_j \psi_j(x)$ adopted for the linear elliptic model with $d := 2^{2(L+1)} - 1$. We consider the general form

$$\beta_l := c2^{-\gamma l}, \;\; c := 0.3\frac{1-2^{-\gamma}}{1-2^{-(L+1)\gamma}} = 0.3\frac{2^\gamma - 1}{2^\gamma - 2^{-L\gamma}}, \tag{5.5.11}$$

which ensures that the uniform ellipticity assumption $\mathbf{UEA}(r,\tilde{R})$ holds with $r = 0.1$ and $\tilde{R} = 1.9$. The value of the parameter $\gamma > 0$ reflects the decay of the high scale oscillation and therefore the long range correlation in the diffusion field.

In our numerical test, we use the value $\gamma = 3$, which was among those tested in Chapter 3 and we consider the maximal scale levels $L = 1$ and 2 which give parametric dimension $d = 15$ and 63. In order to refine the comparison between AI and LNT, we introduce a third process that builds the interpolation polynomials $\mathcal{I}_{\Lambda_n} u \in V_{\Lambda_n}$ by using the sets Λ_n produced by the LNT algorithm. We refer to this algorithm as LNTI (Largest Neighbour Taylor Interpolation). For both AI and LNTI we use the $\Re$-Leja sequence R.

Several observations may be drawn from the error curves, displayed on Figures 5.5.6 and 5.5.7. We first notice that the error curve of LNTI is above that of LNT, with a more oscillatory behaviour. Since the sets $(\Lambda_n)_{n\geq 1}$ are the same for both algorithms, this means that the deterioration is due to the instabilities in the interpolation operator $\mathcal{I}_{\Lambda_n}$ that is used in LNTI, that are reflected by the size of the Lebesgue constant. In

addition, we notice that the error curve of AI is above that of LNTI, which means that AI is slightly misled in the adaptive selection of the sets Λ_n which is better performed by LNT. In all cases, we find that these algorithms are rather robust with respect to the growth in the dimension, since they are able to capture the anisotropic feature of the problem reflected by the decay in the weights β_l.

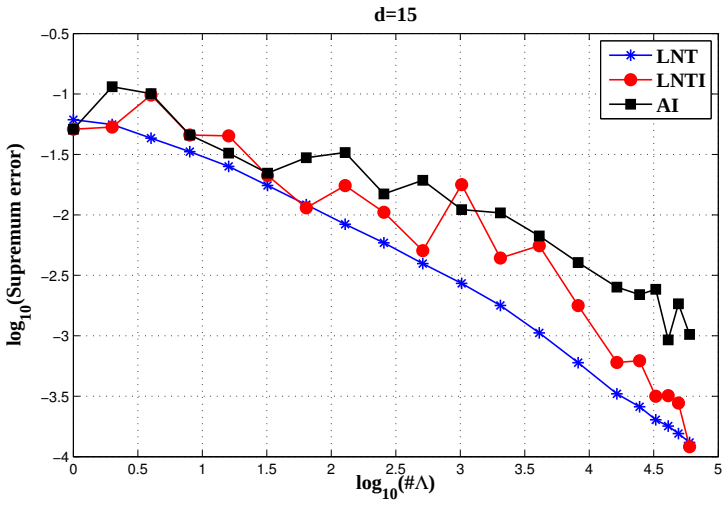

Figure 5.5.6: $\mathcal{V}_\infty$-error of LNT, LNTI and AI for the model (1.1.1) with coefficients (5.5.7) and $d = 15$.

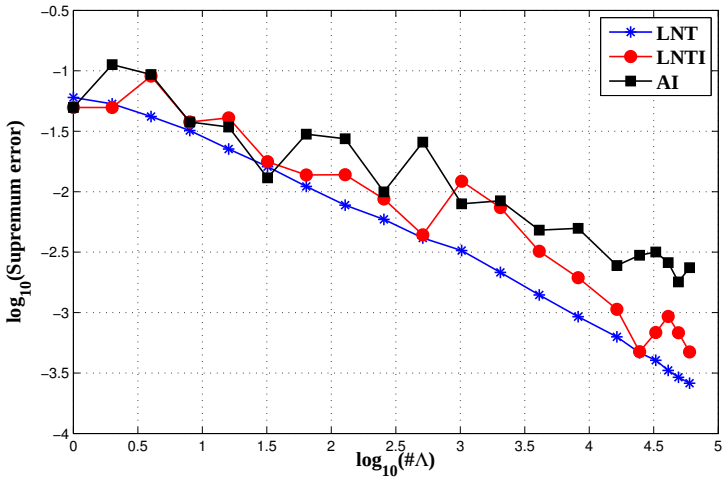

Figure 5.5.7: $\mathcal{V}_\infty$-error of LNT, LNTI and AI for the model (1.1.1) with coefficients (5.5.7) and $d = 63$.

5.6 Extension to non polynomial hierarchical bases

The tensorization-sparsification approach used in the construction of the polynomial interpolation procedure can be generalized to other type of interpolation. We describe the approach in an abstract context using directly the Newton like formula. We consider the following sets

$$(\mathcal{T},\leq), \qquad Z_{\mathcal{T}} := \{z_\lambda \ : \ \lambda\in\mathcal{T}\}, \qquad H_{\mathcal{T}} := \{h_\lambda \ : \ \lambda\in\mathcal{T}\}, \tag{5.6.1}$$

that stands respectively for a countable partially ordered set of indices, a sequence indexed in $\mathcal{T}$ of pairwise distinct abscissas in $[-1,1]$ and a hierarchical basis indexed in $\mathcal{T}$ of functions on $C([-1,1])$ satisfying

$$h_\lambda(z_{\lambda'}) = \delta_{\lambda,\lambda'}, \quad \text{if} \quad \lambda'\leq\lambda. \tag{5.6.2}$$

We consider the set of multi-indices $\mathcal{T}^d := \{\nu=(\nu_1,\dots,\nu_d) : \nu_j\in\mathcal{T}\}$ and define lower sets $\Lambda\subset\mathcal{T}^d$ similarly to (5.1.6) with here $\leq$ is the partial order over $\mathcal{T}$. For a lower set Λ, we introduce

$$\Gamma_\Lambda := \Big\{z_\nu := (z_{\nu_1},\dots,z_{\nu_d}) : \nu\in\Lambda\Big\}, \qquad \mathbb{H}_\Lambda := \operatorname{span}\Big\{H_\nu := \otimes_{j=1}^d h_{\nu_j} : \nu\in\Lambda\Big\}, \tag{5.6.3}$$

the grid of interpolation points and the space of interpolation. The same arguments as used in the proof of Theorem 5.2.1 for the polynomial setting show that the grid Γ_Λ is unisolvant for the space $\mathbb{H}_\Lambda$. In the general context where $\mathcal{T}$ might not be totally ordered, the Smolyak formula (5.2.6) does not make clear sense, yet we may still rely on the recursive computation of the interpolation operators. Namely, if Λ is lower set and $\nu\in\mathcal{T}^d\setminus\Lambda$ such that $\Lambda'=\Lambda\cup\{\nu\}$ is a lower set, then we have

$$\mathcal{I}_{\Lambda'}g = \mathcal{I}_\Lambda g + \Big(g(z_\nu)-\mathcal{I}_\Lambda g(z_\nu)\Big)H_\nu. \tag{5.6.4}$$

Two simple applications for the previous construction are the dyadic hierarchical piecewise linear or piecewise quadratic interpolation. For such interpolation procedures, the set $\mathcal{T}$ is defined by

$$\mathcal{T} = \{\lambda_{-1},\lambda_1,(0,0)\}\cup\Big\{(j,k) \ : \ -2^{j-1}\leq k\leq 2^{j-1}-1, \ j=1,2,\dots\Big\} \tag{5.6.5}$$

induced with the partial order $\lambda_{-1}\leq\lambda_1\leq(0,0)$ and

$$(j,k)\leq(j+1,2k), \qquad (j,k)\leq(j+1,2k+1), \qquad (j,k)\in\mathcal{T}. \tag{5.6.6}$$

The set $\mathcal{T}$ is a binary tree, as represented on Figure 5.6.8, where λ_{-1} is the root node, $(0,0)$ is a child of λ_1 which is a child of λ_{-1}, every node (j,k) has two children $(j+1,2k)$

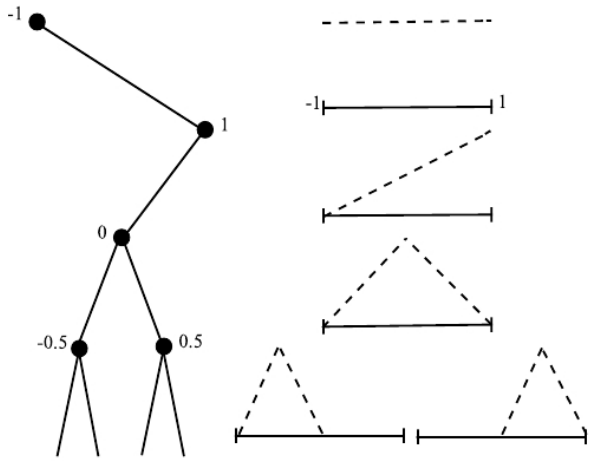

Figure 5.6.8: The binary tree $\mathcal{T}$ and the corresponding univariate piecewise linear hierarchical basis

and $(j+1, 2k+1)$, and the relation $\lambda' \leq \lambda$ means λ' is a parent of λ. We associate with $\mathcal{T}$ the set of abscissas

$$Z_{\mathcal{T}} := \{z_{\lambda_{-1}}, z_{\lambda_1}, z_{(0,0)}\} \cup \Big\{ z_{(j,k)} := \frac{2k+1}{2^j} : (j,k) \in \mathcal{T}, j \geq 1 \Big\}, \tag{5.6.7}$$

where $z_{\lambda_{-1}} = -1$, $z_{\lambda_1} = 1$ and $z_{(0,0)} = 0$. We also associate with $\mathcal{T}$ the hierarchical basis of piecewise linear functions $H_{\mathcal{T}} = \{h_\lambda : \lambda \in \mathcal{T}\}$ defined over $[-1,1]$ by

$$h_{\lambda_{-1}}(s) = 1, \quad h_{\lambda_1}(s) = \frac{1+s}{2}, \quad h_{(j,k)}(x) = \psi(2^j(s - z_{j,k})), \quad \psi(s) := \max\{0, 1-|s|\}, \qquad (j,k) \in \mathcal{T}, \tag{5.6.8}$$

as illustrated on Figure 5.6.8. It is easy to verify that the function h_λ and the abscissas z_λ satisfy the condition (5.6.2), therefore the hierarchical interpolation can be performed. Let us note that in dimension $d = 1$, the hierarchical interpolation amounts in first approximating g by the constant function of value $g(-1)$, second by the affine function that coincides with g at -1 and 1, third by the piecewise affine function that coincides with g at -1, 0 and -1, and in further steps refine by interpolating at the midpoint of an interval between two adjacents interpolation points. The index j corresponds to the level of refinement, or the depth of the node in the binary tree.

In the case of piecewise quadratic interpolation, the procedure is exactly the same except that the hierarchical basis is given by

$$h_{\lambda_{-1}}(s) = 1, \quad h_{\lambda_1}(s) = \frac{(1+s)^2}{4}, \quad h_{(j,k)}(x) = \psi(2^j(s - z_{j,k})), \quad \psi(s) := \max\{0, 1-s^2\}, \qquad (j,k) \in \mathcal{T}. \tag{5.6.9}$$

which corresponds to interpolation by piecewise quadratic functions with the same ordering on the interpolation points as in the piecewise affine case. We should note

that in the two previous cases, the construction for $d = \infty$ is also possible. One considers $\mathcal{F}$ the set of indices in $\nu \in \mathcal{T}^{\mathbb{N}}$ for which only a finite number of indices ν_j are different than λ_{-1} and imitate the procedure given for polynomials.

Having fixed the sparse interpolation procedure (piecewise affine or piecewise quadratic), the objective is now, as in the polynomial case, to select lower sets Λ giving the best possible interpolation spaces $\mathbb{H}_\Lambda$ (or $\mathbb{P}_\Lambda$ in the polynomial case) for the target function g. In the non-intrusive treatment of parametric PDEs that motivated this work, no prior information is known on g, whence the optimal approximation space $\mathbb{H}_\Lambda$ for a given $N = \#(\Lambda)$ is not accessible. Therefore, we may only rely on greedy type strategies such as Algorithm 5.4.4. In view of the recursive formula (5.6.4), we couple the interpolation algorithm with the adaptive strategy for the choice of best multi-index $\nu \in \mathcal{T}^d$ used to enrich Λ. Given Λ a lower set, we denote $\mathcal{N}(\Lambda)$ the set of adjacent neighbours to Λ, which are the multi-indices $\nu \in \mathcal{T}^d \setminus \Lambda$ such that $\Lambda' := \Lambda \cup \{\nu\}$ is a lower set. Depending on the approximation context, we choose to enrich Λ by $\nu \in \mathcal{N}(\Lambda)$ that either satisfies:

- the supremum norm of the increment $\|\mathcal{I}_{\Lambda'} g - \mathcal{I}_\Lambda g\|_{L^\infty(U)}$ is maximal,
- the least square norm of the increment $\|\mathcal{I}_{\Lambda'} g - \mathcal{I}_\Lambda g\|_{L^2(U)}$ is maximal,
- the value $g(z_\nu)$ is maximal.

The two first criterions are designed for the approximation of g in the L^∞ and the L^2 sense, respectively, while the third criterion is designed for optimization and can be seen as a way of exploring the local maxima of g.

Although the set of candidates $\mathcal{N}(\Lambda)$ might be very large, the enrichment step requires at most d new evaluations of g. Indeed, for $\nu \in \Lambda$ and $\Lambda' = \Lambda \cup \{\nu\}$, the set $\mathcal{N}(\Lambda') \setminus \mathcal{N}(\Lambda)$ contains at most d indices. We also note that for the two first strategies, the computation of $\|\mathcal{I}_{\Lambda'} g - \mathcal{I}_\Lambda g\|$ for the new indices consists merely in computing

$$|g(z_\nu) - \mathcal{I}_\Lambda g(z_\nu)| \prod_{j=0}^{d} \|h_{\nu_j}\|, \tag{5.6.10}$$

with $\|\cdot\|$ being the L^∞ or the L^2 norm. This is computationally fast since the h_n are known in advance and their norm $\|h_n\|$ can be tabulated. The cost of the computation $\|\mathcal{I}_{\Lambda'} g - \mathcal{I}_\Lambda g\|$ is essentially dominated by the evaluation of $g(z_\nu)$ in cases where g is evaluated through a heavy numerical solver.

While the above strategies often give good numerical results, they can be defeated when the target function has oscillations that fail to be captured by the greedy selection procedure. For example, consider the two first criterions for piecewise linear adaptive interpolation of a univariate function g such that $g(\frac{1}{2}) = \frac{1}{2}(g(0) + g(1))$. Then, the increment corresponding to the point $z_{(1,0)} = \frac{1}{2}$ is $\Delta_{(1,1)} g = 0$. Therefore the adaptive algorithm might fail to explore the region $[0, 1]$ on which g could still oscillate away

from its linear interpolation. Similarly an algorithm based on the third criterion could be trapped in local maximas.

One way to remedy this defect consists in modifying the algorithm as follows: we produce the nested sequence of lower sets

$$\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \Lambda_n \subset \cdots$$

with $\#(\Lambda_n) = n$, by alternating $p-1$ adaptive steps where the new index $\nu \in \mathcal{N}(\Lambda_n)$ is picked based on the chosen criterion when $n \notin p\mathbb{N}$, and one "conservative" step where we pick the most "ancient" index $\nu \in \mathcal{N}(\Lambda_n)$ when $n \in p\mathbb{N}$, in the sense that it has been lying in $\mathcal{N}(\Lambda_k)$ for the smallest value of $k \leq n$. This conservative step allows us to explore the whole parameter space U, while retaining the adaptive feature of the algorithm. In our numerical test we have used the value $p = 5$, based on empirical observation that it gives a good balance between adaptivity and exploration.

5.7 Conclusion

In this chapter we have introduced an incremental interpolation scheme that can be used in non-intrusive treatment of parametric PDE. The computation of interpolation operators use a Newton like induction formula and its cost, in the framework of parametric PDEs, is mainly dominated by the evaluation of the solution map to be approximated. The scheme can be coupled with an adaptive strategy for the choice of the next interpolation points. The interpolation points lie in a multidimensional grid which is predefined in advance. The stability of the scheme can be controlled if the interpolation grid is obtained by a tensorization of an $\Re$-Leja sequence. Such sequences can have explicit formulas and are studied in the following chapter.

Although, we have not investigated the best choice of points that may yield Lebesgue constant that are logarithmic or linear in the dimension of the polynomials space, we have seen that the choice of a grid based on $\Re$-Leja points provide a cubic Lebesgue constant which can easily absorbed in the convergence rates. This in particular shows that the interpolation scheme presented is well adapted for the parametric PDE studied in chapters 1 and 2.

However, we have only examined the approximation in the uniform sense. The approximation in average sense has not been treated because stability results are not available for such setting. We will propose a non-intrusive least square scheme in Chapter 7 and show that it can yield convergence rates which are near optimal in the average sense.

We have not addressed the convergence of the adaptive algorithm. Numerical experiments suggest that this algorithm yields approximation with near-optimal convergence rates. However, the analysis of its convergence is not clear since the algorithm only

relies on point-wise evaluations of the solution map u and hence does not make full benefit from the smoothness of u.

Chapter 6

Leja sequences on the unit circle and $\Re$-Leja sequences

Contents

6.1 Introduction

In the framework of high dimensional polynomial interpolation presented in the previous chapter, we have seen that the stability of the interpolation process is strongly tied to the stability of the univariate polynomial interpolation process based on the sequence $(z_j)_{j\geq 0}$. More precisely, we have shown that given $Z := (z_i)_{i\geq 0}$ an infinite sequence of mutually disjoint points in $[-1,1]$ (or in a general compact set X of $\mathbb{R}$ or $\mathbb{C}$), with moderate algebraic growth of the Lebesgue constants $\mathbb{L}_{Z_k}$ associated with Lagrange polynomial interpolation at the sections $Z_k := (z_0,\ldots,z_{k-1})$, then the multivariate interpolation processes have also Lebesgue constants with algebraic growth.

For the domain $[-1,1]$ which is of interest to us in the application of the interpolation for the approximation of parametric PDEs, there exists an immense literature dealing with the practical construction of sets of points such that the Lebesgue constant has moderate growth with the number of points k, see the survey [75] and references therein. However, for this domain, the classical sets of optimal k points, for instance Chybeshev abscissas, Gauss-Lobatto abscissas, or more generally the roots of orthogonal classical polynomials, are not nested. This raises the challenge of the construction of infinite sequence $Z \in [-1,1]^{\mathbb{N}}$ with moderate growth of the Lebesgue constants $\mathbb{L}_{Z_k}$.

Solutions for the previously discussed challenge are proposed by means of greedy constructions, such as Leja sequences and magic points [62], and are widely used. Numerical evidence shows that such constructions yield linear growth of the Lebesgue constants, however without any theoretical justification. a simple construction with provable agebraic growth was proposed by Calvi and Phung in [18, 19]. First, in [18], the authors considered $X = \mathcal{U}$ the unit disc in the complex domain and show that for any Leja sequence $E = (e_j)_{j\geq 1}$ with e_0 picked on the boundary $\partial\mathcal{U}$, there exists a constant C such that

$$\mathbb{L}_{E_k} = \mathbb{L}_{\{e_0,\ldots,e_{k-1}\}} \leq Ck\log k, \quad k \geq 2. \tag{6.1.1}$$

Such Leja sequences have a simple geometric structure identified in [17]: when $e_0 = 1$, the section E_{2^n} coincides as a set with the set of the 2^n-roots of unity, and for a more general $e_0 \in \partial\mathcal{U}$ it coincides with the set of 2^n-roots of unity multiplied by e_0.

In addition Calvi and Phung have studied the so-called $\Re$-Leja sequences $R = (r_j)_{j\geq 0}$ obtained by projection of Leja sequences E with $e_0 = 1$ onto the real interval $[-1,1]$, hence taking successively the real parts of the numbers $e_1, e_2, \ldots$, making sure not to project the values e_j for which $e_k = \overline{e}_j$ for some $k < j$, so that the sequence R is of mutually disjoints values. They have proved for the corresponding sections $R_k := (r_0,\ldots,r_{k-1})$ of the sequence R an estimate of the form

$$\mathbb{L}_{R_k} \leq Ck^3\log k, \quad k \geq 2. \tag{6.1.2}$$

Note that, according to the particular structure of Leja sequences on $\mathcal{U}$ when $e_0 = 1$, i.e. $E_{2^{n+1}}$ coincides as a set with the 2^{n+1}-roots of unity, the section R_{2^n+1} coincides as

a set with the so-called Gauss-Lobatto or Clemshaw-Curtis points,

$$\cos(2^{-n}k\pi) \quad \text{for} \quad k = 0, \dots, 2^n \tag{6.1.3}$$

Also note that the sequence R is not a Leja sequence on $[-1, 1]$.

For the same sequences as above, we have established in [21] improved algebraic bounds. Namely

$$\mathbb{L}_{E_k} \leq 2k, \qquad \mathbb{L}_{R_k} \leq 5k^2 \log k, \quad k \geq 2. \tag{6.1.4}$$

These improvements are obtained through a study of structural properties of the Leja sequences E and their projections R. In this chapter, we recall and refine the approach of [21] and establish a new estimate in the real interval case, namely

$$\mathbb{L}_{R_k} \leq 8k^2, \quad k \geq 2. \tag{6.1.5}$$

For this purpose, we exhibit new structural properties of Leja sequences and ℜ-Leja sequences. Exploiting such properties, we investigate the Lebesgue constant δ_k of the difference operator Δ_k, between the interpolation operators associated with nested sections $\{r_0, \dots, r_{k-1}\} \subset \{r_0, \dots, r_k\}$, used in the previous chapter. We have obviously $\delta_k \leq L_{R_{k+1}} + L_{R_k} \leq 16(k+1)^2$. Here we establish the better bound

$$\delta_k \leq (k+1)^2, \tag{6.1.6}$$

which can be used directly, in view of Remark 5.3.2 of the previous chapter, in order to establish that interpolation operators in high dimension based on ℜ-Leja sequences have cubic Lebesgue constant.

In §6.2, we introduce the notations that we adopt for the subsequent sections.

In §6.3, we investigate Leja sequences E on the unit disk $\mathcal{U}$ with starting point in $\partial\mathcal{U}$. We recall their properties as identified in [18] and the simple construction of the so-called simple Leja sequences. Using their definition, we establish, given a sequence E, recursive estimates for the Lebesgue constants $\mathbb{L}_{E_k}$ showing that their growth can be monitored by the value of the Lebesgue function on the next point e_k. Combining this with the particular structure of Leja sequences, we establish the growth bound $2k$.

In §6.5, we describe the novelty of the analysis of this chapter compared to our previous work [21]. First, we give the explicit formula giving an ℜ-Leja sequence $R = (r_j)_{j\geq 1}$ obtained from the projection of E a Leja sequence in $\mathcal{U}$. We then establish a new property of ℜ-Leja sequences stating that $R^2 := (2r_{2j}^2 - 1)_{j\geq 1}$ is also a Leja sequence. Finally, using a simple observation, we show that the analysis of Lebesgue constants $\mathbb{L}_{R_k}$ on $[-1, 1]$ can be implied from the the analysis of Lebesgue constant $\mathbb{L}_{E_{k'}}$ on $\mathcal{U}$ with k' depending on k. This new approach, which is not the one we used in [21], allows us to take benefit from the linear bound obtained in the complex case and recover the bound (6.1.5).

In §6.6, we study the growth of the Lebesgue constant of the difference operators Δ_k both in the complex and real case. Using the structural properties that we established for the complex and real case, we established the bounds $\delta_k \leq k+1$ and $\delta_k \leq (k+1)^2$ respectively.

Finally, in §6.7, we present a numerical illustration of the growth of exact Lebsgue constants associated with Leja sequences E and associated $\Re$-Leja sequences R. We compare the latter with intuitive choicea for sequences with moderate, however not proven, growth of Lebesgue constant.

6.2 Polynomial interpolation on nested sequences

Let $Z_k := \{z_0, \ldots, z_{k-1}\}$ be a set of k pairwise distinct points in a compact set X contained either in $\mathbb{R}$ or $\mathbb{C}$. Any function $f \in C(X)$ admits a unique polynomial interpolant of degree $k-1$ at these points defined by

$$\Pi_{Z_k} f(z) := \sum_{i=0}^{k-1} f(z_i) l_i(z), \tag{6.2.1}$$

where

$$l_j(z) = \prod_{\substack{i=0 \\ i \neq j}}^{k-1} \frac{z - z_i}{z_j - z_i} = \frac{w(z)}{w'(z_j)(z - z_j)} \quad \text{with} \quad w(z) = \prod_{i=0}^{k-1} (z - z_i), \tag{6.2.2}$$

are the associated Lagrange polynomials. The stability of the interpolation process is quantized by the Lebesgue constant

$$\mathbb{L}_{Z_k} := \max_{f \in C(X) - \{0\}} \frac{\|\Pi_{Z_k} f\|_{L^\infty(X)}}{\|f\|_{L^\infty(X)}} = \max_{z \in X} \lambda_{Z_k}(z), \tag{6.2.3}$$

where

$$\lambda_{Z_k}(z) := \sum_{i=0}^{k-1} |l_i(z)|, \tag{6.2.4}$$

is the so-called Lebesgue function.

6.2.1 Leja sequences

Leja sequences on a compact set X are defined by picking an initial point $e_0 \in X$ and defining inductively

$$e_j = \text{Argmax}_{z \in X} |\prod_{l=0}^{j-1} (z - e_l)|. \tag{6.2.5}$$

We mean by (6.2.5) that e_j can be any element in X maximizing the product in the right hand side. Note that such a sequence is in general not uniquely defined since the above maximum can be attained at several points. This procedure may be viewed as a greedy selection that mimics the Fekete points which are defined for a given k as the set of points $\{z_0, \ldots, z_{k-1}\}$ maximizing the product $\prod_{i\neq j} |z_i - z_j|$ over X^k and for which the Lebesgue constant is always smaller than k. We refer to [37] for a survey on Leja sequences. Note that other methods exist to efficiently compute approximate Fekete points [14, 77], however they do not produce the sections of a single sequence, which is our primary motivation. Other greedy approaches that do produce sections of a single sequence have recently been studied in [62]. The points produced by the approaches of [62] are the so-called magic points.

Let us observe that for $k \geq 1$ and Π_{Z_k} the Lagrange interpolation operator associated with $\{z_0, \ldots, z_{k-1}\}$, one has

$$(z - z_0) \ldots (z - z_{k-1}) = z^k - \Pi_{Z_k}(z^k). \tag{6.2.6}$$

Indeed, the real or complex polynomial $z^k - \Pi_{Z_k}(z^k)$ has degree k, has leading coefficient 1 and in view of Lagrange interpolation, has the roots $z_0, \ldots, z_{k-1}$. This shows that for Leja sequences, the next element z_k is chosen among the points where the interpolation error $\|z^k - \Pi_{Z_k}(z^k)\|_{L^\infty(X)}$ is maximal. Moreover, it is easily checked that (6.2.6) is unchanged if one replaces z^k by $z^k + P$ where $P \in P_{k-1}$. This shows that the procedure giving Leja sequences is equivalent to the procedure of the so-called magic points introduced in [62] obtained in X with polynomials spaces of increasing dimension $\mathbb{P}_k$, according to the following construction:

$$w_0 = \text{argmax}_{w\in\mathbb{P}_0} \|w\|_{L^\infty(X)}, \quad z_0 = \text{argmax}_{z\in X} |w_0(z)|, \tag{6.2.7}$$

and $z_0, \ldots, z_{k-1}$ have being constructed and Π_{Z_k} being the polynomial interpolation operator associated with $\{z_0, \ldots, z_{k-1}\}$, then

$$w_k = \text{argmax}_{w\in\mathbb{P}_k} \|w - \Pi_{Z_k}(w)\|_{L^\infty(X)} \quad z_k = \text{argmax}_{z\in X} |w_k(z) - \Pi_{Z_k}(w_k)(z)|. \tag{6.2.8}$$

In the following, given a Leja sequence $E := (e_j)_{j\geq 0}$ on X, we shall call the finite sequences $E_k := (e_0, \ldots, e_{k-1})$ a k-Leja section. For general domains X, there is no theoretical guarantee that Lebesgue constant $\mathbb{L}_{E_k}$ behaves polynomially. We will see in Section 6.3 that the case of complex unit disc $X = \mathcal{U}$ can however be studied.

6.2.2 Binary expansion of integer

The binary representation of integers play a substantial role in the analysis of Leja and ℜ-Leja points. We shall present in details different related notations. Given an

integer $k \geq 1$ and the integer $n \geq 0$ such that $2^n \leq k < 2^{n+1}$, we introduce the binary expansion of k

$$k = a_0 a_1 \dots a_n := \sum_{j=0}^{n} a_j 2^j, \qquad a_j \in \{0, 1\}. \tag{6.2.9}$$

Let us remark that with this definition $a_n = 1$. We denote respectively by $\sigma_1(k)$, $\sigma_0(k)$ and $p(k)$ the number of ones and zeros in the binary expansion of k and the largest integer p such that 2^p divides k, i.e.

$$\sigma_1(k) := \sum_{j=0}^{n} a_j, \qquad \sigma_0(k) := \sum_{j=0}^{n} (1-a_j) = (n+1) - \sigma_1(k), \qquad p(k) := \inf\{0 \leq j \leq n : a_j = 1\}. \tag{6.2.10}$$

In the following sections, we will use in various occasion the induction on binary expansions of integer. For the sake of clarity, we provide here various identities relating the quantities that we just defined. First, for the range of integer $2^n \leq k < 2^{n+1}$ considered, we have

$$\sigma_0(k) = \sigma_1(2^{n+1} - 1 - k), \qquad \sigma_1(k-1) = p(k) + \sigma_1(k) - 1, \tag{6.2.11}$$

with the latter equality valid for any $k \geq 2$. The first equality follows from $\sum_{j=0}^{n}(1 - a_j)2^j = 2^{n+1} - 1 - k$, while the second follows from the observation that for $k = 2^{p(k)}(1 + 2m)$, one has

$$k = \underbrace{00\dots0}_{p(k)} 1m \qquad \text{so that} \qquad k - 1 = \underbrace{11\dots1}_{p(k)} 0m,$$

in the sense of binary expansion. We also have the following identity

$$\sigma_1(k) + \sigma_1(2^n - k) + p(k) = n + 1, \qquad 0 \leq n, \quad 1 \leq k \leq 2^n. \tag{6.2.12}$$

This can be easily checked for $k = 1$. For $k \geq 2$, writing $k = l+1$ and using (6.2.11), we obtain $\sigma_1(k)+p(k) = \sigma_1(l)+1$ and $\sigma_1(2^n-k) = \sigma_1((2^n-l)-1) = \sigma_1(2^n-l)+p(2^n-l)-1$, so that the fact $p(2^n - l) = p(l)$ implies the result for k by induction.

Throughout this chapter, to any finite set S of real or complex numbers, we associate the function

$$w_S(x) := \prod_{s \in S} (x - s). \tag{6.2.13}$$

Given a sequence $E = (e_0, e_1, \dots)$ a finite or infinite sequence of real or complex numbers, we introduce the notations

$$E_{l,m} := (e_l, \dots, e_{m-1}), \qquad \rho E := (\rho e_0, \rho e_1, \dots), \qquad \Re(E) := (\Re(e_0), \Re(e_1), \dots), \tag{6.2.14}$$

where the number l and m are such that $l < m$, ρ a complex or real number and $\Re(s)$ denote the real part of s. Finally, given two finite sequences $A := (a_0, \dots, a_{r-1})$ and $B := (b_0, \dots, b_{s-1})$, we denote by $A \wedge B$ the concatenation of A and B, i.e.

$$A \wedge B := (a_0, \dots, a_{r-1}, b_0, \dots, b_{s-1}) \tag{6.2.15}$$

6.3 Properties of Leja sequences on the unit disk

6.3.1 Construction of Leja sequences

We introduce the notation

$$\mathcal{U}_N := \left\{\rho_N^0, \cdots, \rho_N^{N-1}\right\}, \quad \rho_N = e^{i2\pi/N}, \tag{6.3.1}$$

for the set of the N roots of unity. The structure of Leja sequences on the complex unit disk with initial value 1 is characterized by the following

Theorem 6.3.1 (*Theorem 5 in [17]*)

Let E a Leja sequence on $\mathcal{U}$ with initial value $e_0 = 1$ and $n \geq 0$. The equality $E_{2^n} = \mathcal{U}_{2^n}$ holds in the set sense and for any k such that $2^n < k < 2^{n+1}$, the sequence $U_{k-2^n} := e_{2^n}^{-1} E_{2^n,k} = \frac{1}{e_{2^n}}(e_{2^n}, \ldots, e_{k-1})$ is a $\{k-2^n\}$-Leja section starting at 1.

The previous theorem already shows the strong dependence of Leja sequences on the binary expansion of integers. Let us remark that the previous theorem also holds in the case where $e_0 = \rho \in \partial\mathcal{U}$ general, yet with the slight difference $E_{2^n} = \rho\mathcal{U}_{2^n}$ in the set sense. This implies in particular that for E_k a k-Leja section with initial value $\rho \in \partial\mathcal{U}$, and n such that $2^n \leq k < 2^{n+1}$, one has

$$w_{E_k}(z) = (z^{2^n} - \rho^{2^n}) w_{E_{2^n,k}}(z), \qquad z \in \mathcal{U}, \tag{6.3.2}$$

with the convention $w_{E_{2^n,k}}$ is the constant polynomial 1 when $k = 2^n$. In the case $k > 2^n$, the section $E_{2^n,k}$ is $\{k-2^n\}$-Leja section with initial value $e_{2^n} \in \partial\mathcal{U}$ and one can start over and write $w_{E_{2^n,k}}(z)$ as the previous product. This inductive process was used in [18] in order to give a simple formula of w_{E_k} as a product of $\sigma_1(k)$ translated monomials and to prove that the maximum of $|w_{E_k}|$ over $\mathcal{U}$ is equal to $2^{\sigma_1(k)}$. We use the same approach to prove the same results, yet with a more intuitive form of the polynomial w_{E_k}. We have the following

Lemma 6.3.2

Let $k \geq 1$ with binary expansion (6.2.9), E_k a k-Leja section in $\mathcal{U}$ with initial value $\rho \in \partial\mathcal{U}$ and $e_k \in \partial\mathcal{U}$ maximize $|w_{E_k}|$. Then with $k = \sum_{j=1}^n a_j 2^n$, we have

$$w_{E_k}(z) = \prod_{\substack{0 \leq j \leq n \\ a_j = 1}} (z^{2^j} + e_k^{2^j}), \qquad z \in \mathcal{U}. \tag{6.3.3}$$

In particular, $\sup_{z\in\mathcal{U}} |w_{E_k}(z)| = |w_{E_k}(e_k)| = 2^{\sigma_1(k)}$.

Proof: Let E be is any Leja sequence on $\mathcal{U}$ whose k-section coincides with E_k and $(k+1)$-th element is e_k. By the implications of Theorem 6.3.1, $E_{2^n,2^{n+1}} = \rho(\mathcal{U}_{2^{n+1}} \setminus \mathcal{U}_{2^n})$ holds

in the set sense, so that $E_{2^n,2^{n+1}}$ coincides as a set with the 2^n-roots of $-\rho$. Since $2^n \le k \le 2^{n+1}-1$, this implies that $e_k^{2^n} = -\rho^{2^n}$. We may then rewrite (6.3.2) as

$$w_{E_k}(z) = (z^{2^n} + e_k^{2^n}) w_{E_{2^n,k}}(z), \qquad z \in \mathcal{U}.$$

We observe then that an induction on $k \ge 1$ imposes to us. First, since $e_1 = -\rho$, then $w_{E_1} = (z-\rho) = (z+e_j)$ and the result holds for $k=1$. If $k = 2^n$, then the previous equality is exactly (6.3.3) since $w_{E_{2^n,k}} = 1$. We assume now that $2^n < k < 2^{n+1}$. From the previous equality e_k maximize $w_{E_{2^n,k}}$. Moreover by Theorem 6.3.1, $E_{2^n,k}$ is a l-Leja section with $l = k - 2^n = \sum_{j=0}^{n-1} a_j 2^j$, therefore (6.3.3) for E_k follows from the induction hypothesis applied with $E_{2^n,k}$. The second conclusion of the lemma is a straightforward application of the first one. ■

We should stress that Theorem 6.3.1 completely determines the structure of Leja sequences on the unit disk with initial value in $\partial\mathcal{U}$. The subsequent results on such sequences are merely implication of this structural theorem. We shall also note that the converse of Theorem 6.3.1 holds. It can be stated as follows

Lemma 6.3.3

Let $n \ge 0$, $2^n < k \le 2^{n+1}$ and $l = k - 2^n$. If $E_{2^n} = (e_j)_{0\le j\le 2^n-1}$ and $U_l = (u_j)_{0\le j\le l-1}$ are respectively a 2^n and l-Leja sections starting at 1 and ρ is a 2^n-root of -1, then $E_k = E_{2^n} \wedge \rho U_l$ is a k-Leja section.

Proof: First, by the assumptions of the lemma, for $j = 1, \dots, 2^n - 1$, e_j maximizes $|w_{E_j}|$. Now, by Theorem 6.3.1, $E_{2^n} = \mathcal{U}_{2^n}$ and $U_l \subset \mathcal{U}_{2^n}$ in the set sense. Therefore the assumption $\rho^{2^n} = -1$ implies that for any $j = 2^n, \cdots k-1$, $e_j := \rho u_{j-2^n}$ is a 2^n root of -1, so that it maximizes $|w_{E_{2^n}}(z)| = |z^{2^n} - 1|$. Moreover, if $j \ge 2^n + 1$, then

$$|w_{E_j}(z)| = |z^{2^n} - 1||w_{U_{j-l}}\Big(\frac{z}{\rho}\Big)|, \qquad z \in \mathcal{U},$$

so that since e_j maximizes $|w_{U_{j-l}}(\frac{z}{\rho})|$, then it maximizes $|w_{E_j}(z)|$ as well. ■

The previous lemma and Theorem 6.3.1 shows that the construction of a Leja sequence on $\mathcal{U}$ with initial value in $\partial\mathcal{U}$ amounts to concatenating Leja sections according to a particular rule. The most natural construction consists in defining a sequence $\mathcal{E} := (\xi_j)_{j\ge 0}$ inductively by

$$\mathcal{E}_1 := (e_0 = 1) \quad \text{and} \quad \mathcal{E}_{2^{n+1}} := \mathcal{E}_{2^n} \wedge e^{\frac{i\pi}{2^n}} \mathcal{E}_{2^n}, \quad n \ge 0. \tag{6.3.4}$$

This very uniform pattern of the sequence $\mathcal{E}$ yields an interesting distribution of its elements, see Figure 6.3.1. Indeed, by an immediate induction, see [17], it can be shown that the elements ξ_k are given by

$$\xi_k = \exp\Big(i\pi \sum_{l=0}^{n} a_j 2^{-j}\Big) \text{ for } k = \sum_{j=0}^{s} a_j 2^j, \quad a_j \in \{0,1\}. \tag{6.3.5}$$

The construction yields then a Low-discrepancy sequence on $\partial\mathcal{U}$ based on Van der Corput enumeration. This sequence was known to be a Leja sequence over $\mathcal{U}$ in many earlier works.

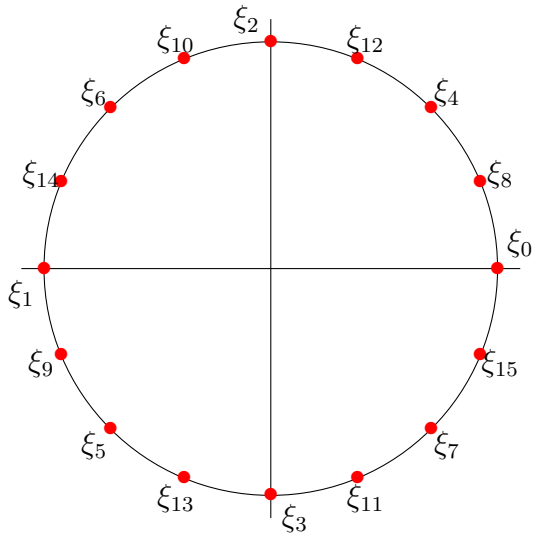

Figure 6.3.1: Distribution of the first 16 elements of the Leja sequence $\mathcal{E}$.

6.3.2 Symmetry properties of Leja sequences on the unit disk

In this paragraph, we give three symmetry properties inherited from the binary pattern of the distribution of Leja sequences on $\mathcal{U}$. The three result are very helpful in the analysis of the Lebesgue constants on the complex unit disk addressed in Section 6.4.2 and the analysis of the Lebesgue constants on the real interval addressed in Section 6.5.

Lemma 6.3.4

Let $E = (e_j)_{j\geq 0}$ be a Leja sequence on $\mathcal{U}$ with initial value in $\partial\mathcal{U}$. The sequences $E^{-1} := (e_j^{-1})_{j\geq 0}$ and $E^2 := (e_{2j}^2)_{j\geq 0}$ are also Leja sequences on $\mathcal{U}$.

Proof: Since the elements of E lie all in $\partial\mathcal{U}$, the sequence E^{-1} is symmetric to E with respect to the abscissas axis. As far as the distances are concerned, E^{-1} and E play symmetric roles, therefore E^{-1} is a Leja sequence. For the second sequence, by Theorem 6.3.1, E satisfies $e_{2j+1} = -e_{2j}$ for any $j \geq 0$, therefore, for $k \geq 1$, one has

$$w_{E_k^2}(z^2) = \prod_{j=0}^{k-1}(z^2 - e_{2j}^2) = \prod_{l=0}^{2k-1}(z - e_l) = w_{E_{2k}}(z). \tag{6.3.6}$$

Consequently, e_{2k} maximizes $|w_{E_k^2}(z^2)|$, which is equivalent to e_{2k}^2 maximizes $|w_{E_k^2}|$. ■

The previous lemma implies immediately that $E^{2^n} := (e^{2^n}_{2^n j})_{j\geq 0}$ is also a Leja sequence over $\mathcal{U}$. We shall use this symmetry property in order to relate the Lebesgue function associated with a $\{2^n k\}$-Leja sections to the Lebesgue function associated a k-Leja section.

The second result is concerned with the comparison of Leja sections with similar length. It states that up to a rotation and a permutation, any two k-Leja section are equal. More precisely, we have the following

Lemma 6.3.5

Let $n \geq 0$, k an integer with $2^n \leq k < 2^{n+1}$ and $E_k := (e_k)_{0\leq j\leq k-1}$ and $F_k := (f_k)_{0\leq j\leq k-1}$ two k-Leja sections on $\mathcal{U}$ with $e_0 = f_0 = 1$. There exists ρ a 2^n-roots of unity such that

$$E_k = \rho F_k \tag{6.3.7}$$

where the above equality is meant in the sets sense.

Proof: We use induction on n. When $n = 0$ then $k = 1$ is an equality holds. We suppose now that $n \geq 0$ and that the result holds for any k with $2^n \leq k < 2^{n+1}$. Let k an integer with $2^{n+1} \leq k < 2^{n+2}$ and E_k and F_k two k-Leja sections with $e_0 = f_0 = 1$. According to Theorem 6.3.1, $E_{2^{n+1}} = F_{2^{n+1}} = \mathcal{U}_{2^{n+1}}$ as sets and

$$E_k = E_{2^{n+1}} \wedge \rho_1 U_{k'}, \quad F_k = F_{2^{n+1}} \wedge \rho_2 V_{k'}, \quad k' = k - 2^{n+1}$$

where ρ_1 and ρ_2 are both 2^{n+1}-roots of -1 and $U_{k'}$ and $V_{k'}$ are both two k'-Leja sections starting at 1. By the induction hypothesis, there exists ρ a 2^n-roots of unity such that $U_{k'} = \rho V_{k'}$ as sets, therefore in the set sense

$$E_k = E_{2^{n+1}} \bigcup \rho_1 U_{k'} = \frac{\rho_1\rho}{\rho_2} E_{2^{n+1}} \bigcup \frac{\rho_1\rho}{\rho_2}\rho_2 V_{k'} = \frac{\rho_1\rho}{\rho_2}(F_{2^{n+1}} \bigcup \rho_2 V_{k'}) = \frac{\rho_1\rho}{\rho_2} F_k.$$

We have used the fact that $E_{2^{n+1}} = \frac{\rho_1\rho}{\rho_2} E_{2^{n+1}}$ as sets which follows from $E_{2^{n+1}} = \mathcal{U}_{2^{n+1}}$ and $(\frac{\rho_1\rho}{\rho_2})^{2^{n+1}} = 1$. ∎

The previous result is obviously true for general k-Leja sections with initial values in $\partial\mathcal{U}$, yet with the difference $\rho \in \partial U$ will not be necessarily a 2^n-root of 1. Since the Lebesgue constants associated with points in $\mathcal{U}$ are invariant by rotation of the points, then the Lebesgue constants of all k-Leja sections E_k with $e_0 \in \partial\mathcal{U}$ are equal and only depends on k. We will denote by $\mathbb{L}_k$ the common value.

The last symmetry result is concerned with the structure of Leja sections when enumerated in the backward sense. For a finite sequence $Z_k = (z_j)_{0\leq j\leq k-1}$ of points on $\mathcal{U}$, we introduce the notation

$$\mathcal{B}(Z_k) := (z_{k-1-j})_{0\leq j\leq k-1} \tag{6.3.8}$$

for the finite sequence of points of Z_k with backward indexing. We are interested in the structure of such sections for Leja sequences on $\mathcal{U}$. Consider the sequence $\mathcal{E}$

defined in (6.3.5) and $n \geq 0$. We claim that $\mathcal{B}(\mathcal{E}_{2^n})$ is a 2^n-Leja section. Indeed, Given $j = \sum_{l=0}^{n-1} a_l 2^l \in \{0, \cdots, 2^n - 1\}$, one has $2^n - 1 - j = \sum_{l=0}^{n-1}(1 - a_l)2^l$, so that

$$\xi_{2^n-1-j} = \exp\Big(i\pi \sum_{l=0}^{n-1}(1 - a_l)2^{-l}\Big) = \frac{\rho_n}{\xi_j}, \qquad \rho_n := e^{-\frac{i\pi}{2^{n-1}}}. \tag{6.3.9}$$

Therefore, according to Lemma 6.3.4, $\mathcal{B}(\mathcal{E}_{2^n}) = \rho_n \mathcal{E}_{2^n}^{-1}$ is indeed a 2^n-Leja section. In general, the following result also holds.

Lemma 6.3.6

For any $n \geq 0$ and any 2^n-Leja section $E_{2^n} = (e_j)_{0 \leq j \leq 2^n-1}$ with $e_0 \in \partial\mathcal{U}$, $\mathcal{B}(E_{2^n})$ is also a 2^n-Leja section on $\mathcal{U}$.

Proof: Up to rotate $F_{2^n} = \mathcal{B}(E_{2^n})$, we may suppose that $f_0 = 1$ and use induction on $n \geq 0$. It is obvious for $n = 0$, and we assume it is true for a $n \geq 0$. We consider $E_{2^{n+1}}$ a 2^{n+1}-Leja section with $e_0 = 1$. By Theorem 6.3.1, E_{2^n} and $E_{2^n,2^{n+1}}$ are 2^n-Leja sections, therefore by the induction hypothesis, so are $\mathcal{B}(E_{2^n})$ and $\mathcal{B}(E_{2^n,2^{n+1}})$. Moreover, with $\rho_n = e_{2^n-1}$ and $\rho_{n+1} = e_{2^{n+1}-1}$, we have $\frac{1}{\rho_{n+1}}\mathcal{B}(E_{2^n,2^{n+1}})$ and $\frac{1}{\rho_n}\mathcal{B}(E_{2^n})$ are 2^n-Leja section initiated at 1 and

$$\mathcal{B}(E_{2^{n+1}}) = \mathcal{B}(E_{2^n,2^{n+1}}) \wedge \mathcal{B}(E_{2^n}) = \rho_{n+1}\Big(\frac{1}{\rho_{n+1}}\mathcal{B}(E_{2^n,2^{n+1}}) \wedge \rho \frac{1}{\rho_n}\mathcal{B}(E_{2^n})\Big), \qquad \rho = \frac{\rho_n}{\rho_{n+1}}.$$

By theorem 6.3.1, ρ_n and ρ_n are 2^n-root of 1 and -1 respectively, therefore ρ is a 2^n-root of -1. Applying finally Lemma 6.3.3, we deduce that $\mathcal{B}(E_{2^{n+1}})$ which completes the proof. ■

The previous lemma has an implication on the minimal growth of the polynomials w_{E_k} for k-Leja sections E_k that turn out to be useful in the analysis of the growth of Lebesgue constant in the real and complex case. We have the following

Corollary 6.3.7

Let $n \geq 0$, $1 \leq k \leq 2^n$ and E_k be a k-Leja section on $\mathcal{U}$ with initial value $\rho \in \partial\mathcal{U}$. For any $\xi \in \mathcal{U} \setminus E_k$

$$\frac{1}{|w_{E_k}(\xi)|} \leq \frac{2^{\sigma_1(l)}}{|\xi^{2^n} - \rho^{2^n}|}, \qquad l = 2^n - k, \tag{6.3.10}$$

with the convention $\sigma_1(0) = 1$.

Proof: Let F_l be a finite sequence that completes E_k to a 2^n-Leja section, i.e. $E_{2^n} := E_k \wedge F_l$ is a 2^n-Leja section. By the previous lemma $\mathcal{B}(E_{2^n})$ is a 2^n-Leja section, hence

$\mathcal{B}(F_l)$ is a l-Leja section. In view of Lemma 6.3.2 and the fact $E_{2^n} = \rho\mathcal{U}_{2^n}$ in the set sense, we deduce that

$$\frac{1}{|w_{E_k}(\xi)|} = \frac{|w_{F_l}(\xi)|}{|w_{E_{2^n}}(\xi)|} = \frac{|w_{\mathcal{B}(F_l)}(\xi)|}{|\xi^{2^n} - \rho^{2^n}|} \leq \frac{2^{\sigma_1(l)}}{|\xi^{2^n} - \rho^{2^n}|}.$$

which is also valid for the case $k = 2^n$, in which case $l = 0$. ■

6.4 Lebesgue constant of Leja sequences on $\mathcal{U}$.

We have proved many interesting properties of Leja sequences on the unit disk. We now are able to give the result on the growth of Lebesgue constant.

6.4.1 Logarithmic estimates

We consider $n \geq 0$, $2^n \leq k < 2^{n+1}$ and E_k a k-Leja section starting at 1. Since $E_k = E_{2^n} \wedge E_{2^n,k}$, then It is easily checked that

$$\mathbb{L}_{E_k} \leq Q(E_{2^n,k}, E_{2^n})\mathbb{L}_{E_{2^n}} + Q(E_{2^n}, E_{2^n,k})\mathbb{L}_{E_{2^n,k}} \quad \text{with} \quad Q(A,B) = \sup_{\substack{z\in\mathcal{U}\\ \xi\in B}} \frac{|w_A(z)|}{|w_A(\xi)|} \tag{6.4.1}$$

Since $E_{2^n} = \mathcal{U}_{2^n}$ in the set sense and $E_{2^n,k}$ is a subset of the set 2^n-root of -1, then $Q(E_{2^n}, E_{2^n,k}) = 1$. In addition Corollary 6.3.7 applied with $E_{2^n,k}$ that is a k'-Leja section with $k' = k - 2^n$ and $\rho = e_{2^n}$ that is 2^n-root of -1 implies in view of (6.2.12)

$$Q(E_{2^n,k}, E_{2^n}) \leq \frac{2^{\sigma_1(k')}2^{\sigma_1(2^n-k')}}{2} = 2^{n-p(k')} = 2^{n-p(k)} \tag{6.4.2}$$

Since the Lebesque constant of k-Leja section with initial value in $\partial\mathcal{U}$ only depends on k, we may rewrite (6.4.1) as

$$\mathbb{L}_k \leq 2^{n-p(k)}\mathbb{L}_{\mathcal{U}_{2^n}} + \mathbb{L}_{k-2^n} \tag{6.4.3}$$

An immediate induction shows that for $k = \sum_{j=p(k)}^{n} a_j 2^j$

$$\mathbb{L}_k \leq \sum_{\substack{j=p(k)\\ a_j=1}}^{n} 2^{j-p(k)}\mathbb{L}_{\mathcal{U}_{2^j}} \tag{6.4.4}$$

Knowing the growth of the Lebesgue constant $\mathbb{L}_{\mathcal{U}_{2^j}}$ from [55] or [21], we infer

$$\mathbb{L}_k \leq \frac{k}{2^{p_0(k)}}\frac{2}{\pi}\Big(\frac{9}{4} + \log 2^n\Big), \tag{6.4.5}$$

Let us remark that this formula is also valid for the case $k = 2^n$ since $p(k) = n$ in this case.

6.4.2 Linear estimates

The bound (6.4.5) is asymptotically sharp for certain values of k, for instance $k = 2^n$. However, numerical evidence shows that (6.4.5) is a pessimistic bound. For example if k is an odd integer, (6.4.5) only gives $\mathbb{L}_k \lesssim k \log k$ while numerical computations indicate that $\mathbb{L}_k \leq k$. In [18] it was conjectured that $\mathbb{L}_k \leq k$. We have shown in [21] that $\mathbb{L}_k \leq 2k$ and therefore that the Lebesgue constant grows at worse linearly. We shall give the steps of the proof of the result as in [21] yet with simpler arguments. We begin first by a recursive result on the the growth of Lebesgue functions valid for Leja sequences on any real or complex domain X.

Theorem 6.4.1

Let E be a Leja sequence on a real or complex compact X. For any $k \geq 1$ and any $z \in X$, it holds

$$\lambda_{E_{k+1}}(z) \leq \lambda_{E_k}(z) + \Big(\lambda_{E_k}(e_k) + 1\Big). \tag{6.4.6}$$

In particular $\mathbb{L}_{E_{k+1}} \leq 2\mathbb{L}_{E_k} + 1$. Moreover

$$\lambda_{E_k}(z) \leq \lambda_{E_{k+1}}(z) + \Big(\lambda_{E_k}(e_k) - 1\Big). \tag{6.4.7}$$

Proof: We fix $k \geq 1$ and denote $l_0, \ldots, l_{k-1}$ the Lagrange polynomials associated with the section E_k and $L_0, \ldots, L_k$ the Lagrange polynomials associated with the section E_{k+1}. By Lagrange interpolation formula, for $j = 0, \ldots, k-1$

$$l_j(z) = \sum_{i=0}^{k} l_j(e_i) L_i(z) = L_j(z) + l_j(e_k) L_k(z)$$

hence

$$\Big| |L_j(z)| - |l_j(z)| \Big| \leq |L_j(z) - l_j(z)| \leq |l_j(e_k)||L_k(z)|$$

The summation over all $j \in \{0, \ldots, k-1\}$ implies

$$\Big| \lambda_{E_{k+1}}(z) - |L_k(z)| - \lambda_{E_k}(z) \Big| \leq \lambda_{E_k}(e_k)|L_k(z)|$$

Using the definition of Lagrange polynomials, $L_k(z) = w_{E_k}(z)/w_{E_k}(e_k)$, we observe that the Leja defnition (6.2.5) implies $|L_k(z)| \leq 1$ for any $z \in X$. Moreover, since any Lebesgue function has a minimum value equal to 1, then $\lambda_{E_k}(e_k) \geq 1$. The previous inequality implies then the two inequalities of the Theorem. ■

The previous theorem shows that the growth of the Lebesgue constants of sections of a Leja sequence E is strongly tied to the growth of the quantities $\lambda_{E_k}(e_k)$. For instance,

it can be easily shown by induction on k that the inequality (6.4.6) alone implies the following

$$\lambda_k(e_k) = \mathcal{O}(\log(k)) \Longrightarrow \mathbb{L}_k = \mathcal{O}(k\log(k)), \tag{6.4.8}$$

and

$$\lambda_k(e_k) = \mathcal{O}(k^\theta) \Longrightarrow \mathbb{L}_k = \mathcal{O}(k^{\theta+1}). \tag{6.4.9}$$

For Leja sequences on the unit disk with initial value in $\partial\mathcal{U}$, we shall show that $\lambda_k(e_k) \leq k$ for any $k \geq 1$, which implies $\mathbb{L}_k = \mathcal{O}(k^2)$. In order to sharpen this bound to $2k$, we should make use of the particular structure of Leja sequences on the unit disk.

For the sake of clarity, we only work with the particular Leja sequence $\mathcal{E}$ defined in (6.3.5). The subsequents results can be stated for general Leja sequence, however, this is irrelevant in our analysis for the growth of Lebesgue constants since they only depends on k. Let us remark that the definition (6.3.5) induces on $\mathcal{E} = (\xi_j)_{j\geq 0}$ the additional symmetry property

$$\mathcal{E}^2 = \mathcal{E}, \quad \text{i.e.} \quad \xi_{2j}^2 = \xi_j, \quad j \geq 0. \tag{6.4.10}$$

The binary patten of the sequence $\mathcal{E}$ yields the following first result.

Lemma 6.4.2

For any $k \geq 1$ and $n \geq 0$, one has

$$\lambda_{\mathcal{E}_{2^n k}}(z) \leq \mathbb{L}_{2^n}\lambda_{\mathcal{E}_k}(z^{2^n}), \quad z \in \mathcal{U}. \tag{6.4.11}$$

In particular $\mathbb{L}_{2k} \leq \sqrt{2}\mathbb{L}_k$.

Proof: By the particular structure of Leja sequences on the unit disk characterized in Theorem 6.3.1, for any $0 \leq j \leq k-1$, $\mathcal{E}_{2^n j, 2^n(j+1)} = (\xi_{2^n j}, \cdots, \xi_{2^n(j+1)-1})$ is a 2^n-Leja section, therefore $\mathcal{E}_{2^n j, 2^n(j+1)} = \xi_{2^n j}\,\mathcal{U}_{2^n}$ in the set sense. This yields, for any $z \in \partial\mathcal{U}$

$$w_{\mathcal{E}_{2^n k}}(z) = \prod_{j=0}^{k-1} w_{\mathcal{E}_{2^n j, 2^n(j+1)}}(z) = \prod_{j=0}^{k-1}(z^{2^n} - \xi_{2^n j}^{2^n}) = \prod_{j=0}^{k-1}(z^{2^n} - \xi_j) = w_{\mathcal{E}_k}(z^{2^n}). \tag{6.4.12}$$

This implies that $w'_{\mathcal{E}_{2^n k}}(z) = 2^n z^{2^n-1} w'_{\mathcal{E}_k}(z^{2^n})$, so that for any $j = 0, \cdots k-1$ and any $l = 0, \cdots, 2^n - 1$

$$|w'_{\mathcal{E}_{2^n k}}(\xi_{2^n j + l})| = 2^n |w'_{\mathcal{E}_k}(\xi_{2^n j}^{2^n})| = 2^n |w'_{\mathcal{E}_k}(\xi_j)|. \tag{6.4.13}$$

We denote by $l_0, \ldots, l_{2^n k - 1}$ the Lagrange polynomials associated with the section $\mathcal{E}_{2^n k}$ and by $L_0, \ldots, L_{k-1}$ the Lagrange polynomials associated with the section $\mathcal{E}_k$. Using the two previous equalities, we infer that

$$|l_{2^n j + l}(z)| = L_j(z^{2^n}) \frac{|z^{2^n} - \xi_j|}{2^n |z - \xi_{2^n j + l}|} = L_j(z^{2^n}) \frac{|z^{2^n} - \xi_{2^n j}^{2^n}|}{2^n |z - \xi_{2^n j + l}|}$$

Since as we have already stated $B_j := \mathcal{E}_{2^n j, 2^n(j+1)} = \xi_{2^n j}\, \mathcal{U}_{2^n}$, then

$$\sum_{l=0}^{2^n-1} |l_{2^n j+l}(z)| = |L_j(z^{2^n})| \sum_{l=0}^{2^n-1} \frac{|z^{2^n} - \xi_{2^n j}^{2^n}|}{2^n |z - \xi_{2^n j+l}|} = |L_j(z^{2^n})| \lambda_{B_j}(z) \leq |L_j(z^{2^n})| \mathbb{L}_{B_j} = |L_j(z^{2^n})| \mathbb{L}_{2^n}.$$

The summation over $j = 0, \ldots, k-1$ implies the desired inequality. The result shows in particular that $\mathbb{L}_{\mathcal{E}_{2k}} \leq \mathbb{L}_2 \mathbb{L}_{\mathcal{E}_k}$, so that the elementary calculation $\mathbb{L}_2 = \mathbb{L}_{(-1,1)} = \sqrt{2}$ completes the proof of the lemma. ■

We now turn to the analysis of the Lebesgue function $\lambda_{\mathcal{E}_k}$ in the case where k is an odd integer. For the needs of our purpose, we only focus on the analysis of the growth of the quantities $\lambda_{\mathcal{E}_k}(\xi_k)$. We have the following

Lemma 6.4.3

For any $k \geq 1$, we have

$$\lambda_{\mathcal{E}_k}(\xi_k) \leq k \tag{6.4.14}$$

Proof: First, we remark that Lemma 6.4.2 implies

$$\lambda_{\mathcal{E}_{2k}}(\xi_{2k}) \leq \sqrt{2} \lambda_{\mathcal{E}_k}(\xi_{2k}^2) = \sqrt{2} \lambda_{\mathcal{E}_k}(\xi_k).$$

Therefore, it is sufficient to prove (6.4.14) for k odd which we now assume and write $k = 2N+1$ with $N \geq 1$. We have

$$w_{\mathcal{E}_k}(z) = (z - \xi_{2N}) w_{\mathcal{E}_{2N}}(z) = (z - \xi_{2N}) w_{\mathcal{E}_N}(z^2),$$

therefore taking $\xi_k = \xi_{2N+1} = -\xi_{2N}$, one infers

$$|w_{\mathcal{E}_k}(\xi_k)| = 2 |w_{\mathcal{E}_N}(\xi_N)| \tag{6.4.15}$$

The derivation of $w_{\mathcal{E}_k}$ with respect to z and the evaluation at ξ_{2j} and ξ_{2j+1} for $j = 0, \cdots, N-1$ yields

$$|w'_{\mathcal{E}_k}(\xi_{2j})| = 2|\xi_{2j} - \xi_{2N}||w'_{\mathcal{E}_N}(\xi_j)| \quad \text{and} \quad |w'_{\mathcal{E}_k}(\xi_{2j+1})| = 2|\xi_{2j+1} - \xi_{2N}||w'_{\mathcal{E}_N}(\xi_j)|$$

Again, since $\xi_k = \xi_{2N+1} = -\xi_{2N}$ and $\xi_{2j}^2 = \xi_{2j+1}^2 = \xi_j$, we deduce

$$|\xi_k - \xi_{2j}||w'_{\mathcal{E}_k}(\xi_{2j})| = 2|\xi_N - \xi_j||w'_{\mathcal{E}_N}(\xi_j)| \quad \text{and} \quad |\xi_k - \xi_{2j+1}||w'_{\mathcal{E}_k}(\xi_{2j+1})| = 2|\xi_N - \xi_j||w'_{\mathcal{E}_N}(\xi_j)| \tag{6.4.16}$$

If we denote by $l_0, \ldots, l_{2N}$ the Lagrange polynomials associated with $\mathcal{E}_k$ and by $L_0, \ldots, L_{N-1}$ the Lagrange polynomials associated with $\mathcal{E}_N$, then by (6.4.15) and (6.4.16), we have

$$|l_{2j}(\xi_k)| = |l_{2j+1}(\xi_k)| = |L_j(\xi_N)|, \qquad j = 0, \ldots, N-1.$$

Combining this with

$$L_{2N}(\xi_k) = \frac{w_{\mathcal{E}_{2N}}(\xi_k)}{w_{\mathcal{E}_{2N}}(\xi_{2N})} = \frac{w_{\mathcal{E}_N}(\xi_k^2)}{w_{\mathcal{E}_N}(\xi_N)} = 1,$$

we deduce that the Lebesgue functions associated with $\mathcal{E}_k$ and $\mathcal{E}_N$ satisfies

$$\lambda_{\mathcal{E}_k}(\xi_k) = 2\lambda_{\mathcal{E}_N}(\xi_N) + 1. \tag{6.4.17}$$

From this, the proof can be completed using induction on k. ■

Let us note that an induction on $n \geq 0$ shows that an equality $\lambda_{\mathcal{E}_k}(\xi_k) = k$ holds for the values of type $k = 2^n - 1, n \geq 0$. This was proved in [18, Theorem 9] and shows that the growth of the Lebesgue constant is at worse linear for such values of k. In the following, we give the main result concerning the growth of Lebesgue constant for Leja sequences on the unit circle.

Theorem 6.4.4

For any $k \geq 1$, we have

$$\mathbb{L}_k = \mathbb{L}_{\mathcal{E}_k} \leq 2k, \tag{6.4.18}$$

which yields

$$\mathbb{L}_k \leq 2\frac{\mathbb{L}_{2^{p(k)}}}{2^{p(k)}}k \tag{6.4.19}$$

Proof: In view of (6.4.2), the second equality follows immediately from the first one. To prove (6.4.18) We use induction on k. The result is true for $k = 1, 2, 3$, since direct computation shows that $\mathbb{L}_k \leq k$ for these values. Now we assume the bound (6.4.18) true for any $j < 4k$, then the induction hypothesis combined with the inequalities $\mathbb{L}_{k+1} \leq 2\mathbb{L}_k + 1$ and $\mathbb{L}_{2k} \leq \sqrt{2}\mathbb{L}_k$ given in Lemmas 6.4.1 and 6.4.2 implies

$$\mathbb{L}_{4k} \leq 2\mathbb{L}_k \leq 4k \leq 8k,$$

$$\mathbb{L}_{4k+1} \leq 2\mathbb{L}_{4k} + 1 \leq 4\mathbb{L}_k + 1 \leq 8k + 1 \leq 2(4k + 1),$$

and

$$\mathbb{L}_{4k+2} \leq \sqrt{2}\mathbb{L}_{2k+1} \leq \sqrt{2}(2\mathbb{L}_{2k} + 1) \leq 4\mathbb{L}_k + \sqrt{2} \leq 8k + 4 = 2(4k + 2).$$

In addition using (6.4.7) and (6.4.14), we deduce

$$\mathbb{L}_{4k+3} \leq \mathbb{L}_{4k+4} + \lambda_{\mathcal{E}_{4k+3}}(\mathcal{E}_{4k+3}) - 1 = \mathbb{L}_{4k+4} + 4k + 2 \leq 2\mathbb{L}_{k+1} + 4k + 2 \leq 4(k+1) + 4k + 2 = 2(4k+3). \tag{6.4.20}$$

Therefore (6.4.18) holds for any $j < 4(k + 1)$ which completes the induction and the proof. ■

The bound 6.4.19 is not sharp since it implies the bound $\mathbb{L}_k \leq 2k$ while numerical computation (figure 6.7) shows that $\mathbb{L}_k$ is much smaller. We conjecture that given E a Leja sequence with $e_0 \in \partial\mathcal{U}$, the exact value of the Lebesgue constant is given by

$$\mathbb{L}_{E_k} = \lambda_{E_k}(e_k) \tag{6.4.21}$$

This is already known to be true for the values of type $k = 2^n$ and $k = 2^n - 1$. Indeed, if $e_0 = 1$, then $E_{2^n} = \mathcal{U}_{2^n}$ and $E_{2^n-1} = \mathcal{U}_{2^n} \setminus \{e_{2^n}\}$ and it was shown that in general the Lebesgue fonction associated with $\mathcal{U}_N$ attains its maximum at the N-roots of -1 and that the Lebesgue fonction associated with $\mathcal{U}_N \setminus \{\rho\}$ with $\rho^N = 1$ attains its maximum at ρ, see [21] and [18] for justifications. For more general value of k, numerical evidence shows that the conjecture seems to be true. In addition, given E a Leja sequence with $e_0 = 1$ and $l_0, \ldots, l_{k-1}$ the Lagrange polynomials associated with the section E_k, it seems that e_k maximizes both $\sum_{j=0}^{2^n-1} |l_j|$ and $\sum_{j=2^n}^{k-1} |l_j|$. Since

$$\sum_{j=2^n}^{k-1} |l_j| = \frac{|z^{2^n-1}|}{2} \lambda_{E_{2^n,k}}(z) \tag{6.4.22}$$

and $E_{2^n,k}$ is a Leja section, then since $e_k^{2^n} = -1$ assuming that $\lambda_{E_{2^n,k}}(z)$ attains its maximum at e_k, the sum $\sum_{j=2^n}^{k-1} |l_j|$ will also attains its maximum at e_k and we see that an induction on k combined with the proof that $\sum_{j=0}^{2^n-1} |l_j|$ is maximum at e_k yields a positive answer for the conjecture.

In the following, we provide a partial answer to the conjecture. We show that given E a Leja sequence with $e_0 \in \partial\mathcal{U}$, then λ_{E_k} considered on $\partial\mathcal{U}$ has a local maximum at e_k.

Lemma 6.4.5

Let E be a Leja sequence on $\mathcal{U}$ with $e_0 \in \partial\mathcal{U}$. For $n \geq 0$ and k such that $2^n \leq k < 2^{n+1}$, We have

$$\lambda_{E_k}(z) \leq \lambda_{E_k}(e_k), \qquad z = e^{i\theta}, |\theta - \theta_k| \leq \frac{\pi}{2^n} \tag{6.4.23}$$

where θ_k is the argument of e_k.

Proof: First, we observe that for a given $i = 0, \ldots, k-1$, we have that e_k^2/e_i is also an element of E_k. Indeed, writing $k = \sum_{j=0}^{n} a_j 2^j$ and using Lemma 6.3.2, we have that there exists $j_i \in \{0, \ldots, n\}$ such that $e_i^{2^{j_i}} + e_k^{2^{j_i}} = 0$, hence

$$\left(\frac{e_k^2}{e_i}\right)^{2^{j_i}} + e_k^{2^{j_i}} = \left(\frac{e_k}{e_i}\right)^{2^{j_i}} (e_k^{2^{j_i}} + e_k^{2^{j_i}}) = 0$$

then again by Lemma 6.3.2, e_k^2/e_i is a root of w_{E_k} which is equivalent to $e_k^2/e_i \in E_k$. Now, for i and j_i as before, we have by Lemma 6.3.2,

$$|w'_{E_k}(e_i)| = 2^{j_i} \prod_{\substack{j=0, \\ j \neq j_i, a_j = 1}}^{n} |e_i^{2^j} + e_k^{2^j}| = |w'_{E_k}(e_k^2/e_i)|, \tag{6.4.24}$$

where the second equality is obtained by the same arguments above. We now are able to prove the result of the lemma. We denote $l_0, \ldots, l_{k-1}$ the Lagrange polynomials associated with E_k and for $i = 0, \ldots, k-1$, we denote $\tilde{l}_i$ the Lagrange polynomial

associated with e_k^2/e_i. Using that $z \mapsto e_k^2/z$ is a bijection from E_k into E_k and pairing the polynomials l_i with $\tilde{l}_i$ taking into account (6.4.24), we deduce

$$2\lambda_{E_k}(z) = \sum_{i=0}^{k}\Big(|l_i(z)|+|\tilde{l}_i(z)|\Big) = \sum_{i=0}^{k}\Big(\frac{|z^{2^{j_i}}+e_k^{2^{j_i}}|}{2^{j_i}|z-e_i|}+\frac{|z^{2^{j_i}}+e_k^{2^{j_i}}|}{2^{j_i}|z-e_k^2/e_i|}\Big)\prod_{\substack{j=0,\\ j\neq j_i, a_j=1}}^{n}\frac{|z^{2^j}+e_k^{2^j}|}{|e_i^{2^j}+e_k^{2^j}|} \tag{6.4.25}$$

We claims that every term in the last sum attains it maximum at $z=e_k$ for the range of z considered in the lemma. Indeed, we have

$$\frac{|z^{2^{j_i}}+e_k^{2^{j_i}}|}{2^{j_i}|z-e_i|}+\frac{|z^{2^{j_i}}+e_k^{2^{j_i}}|}{2^{j_i}|z-e_k^2/e_i|} = \frac{|\xi^{2^{j_i}}+1|}{2^{j_i}}\Big(\frac{1}{|\xi-f_i|}+\frac{1}{|\xi-\overline{f_i}|}\Big)$$

where $\xi = z/e_k = e^{i\phi}$, $|\phi| \leq \frac{\pi}{2^n} \leq \frac{\pi}{2^{j_i}}$ and $f_i = e_i/e_k$ satisfies $(f_i)^{2^{j_i}} = -1$. The previous quantity as a function of ξ is invariant by conjugation of ξ, hence can be only considered for $0 \leq \phi \leq \frac{\pi}{2^{j_i}}$. By elementary trigonometric arguments, it has been shown that the previous quantity on ξ considered over $\{\xi = e^{i\phi}, 0 \leq \phi \leq \frac{\pi}{2^{j_i}}\}$ attains its maximum at $\xi = 1$, see Lemma 2.3 in [21] for a proof, hence over $\{z = e^{i\theta}, |\theta-\theta_k| \leq \frac{\pi}{2^n}\}$, the quantity on z is maximal at $z = e_k$. Injecting this back into (6.4.25) and remarking that the products are also maximal at $z = e_k$, we infer $2\lambda_{E_k}(z) \leq 2\lambda_{E_k}(e_k)$, which finishes the proof. ■

6.5 Lebesgue constant of the ℜ-Leja sequences on $[-1,1]$.

In this section, we address the growth of Lebesgue constants of the projection on the real interval $X := [-1,1]$ of Leja sequence on $\mathcal{U}$ starting at $e_0 = 1$.

6.5.1 Construction of ℜ-Leja sequences on $[-1,1]$.

We consider a Leja sequence $E = (e_j)_{j\geq 0}$ on the unit disk with $e_0 = 1$ and project it onto the real interval $[-1,1]$ and denote by $R = (r_j)_{j\geq 0}$ the sequence obtained. Since $E = (1,-1,\pm i,\cdots)$, one should make sure that no point is repeated on R simply by not projecting a point e_j such that $e_j = \overline{e}_i$ for some $i < j$. Such sequence R was named an ℜ-Leja sequence in [19]. The projection rule that prevent the repetition is well understood. Indeed, it was in proved in [19, Theorem 2.4] that

Lemma 6.5.1

Let E be a Leja sequence on $\mathcal{U}$ with $e_0 = 1$ and R the associated ℜ-Leja sequence. Then

$$R = \Re(Z), \quad \text{with} \quad Z := (1,-1) \wedge \bigwedge_{j=1}^{\infty} E_{2^j,2^j+2^{j-1}}. \tag{6.5.1}$$

The previous formula obviously means that R is the projection element-wise of the sequence Z. A straightforward cardinality argument shows that in addition to $r_0 = 1$, $r_1 = -1$, for any $n \geq 0$ and any k with $2^n \leq k - 1 < 2^{n+1}$, r_k is explicitly given by

$$r_k = \Re(e_{2^n+k-1}). \tag{6.5.2}$$

For instance if E is the simple Leja sequence given in (6.3.5), then for any $n \geq 0$ and any k with $2^n \leq k < 2^{n+1}$ having the binary expansion $k = 2^n + \sum_{j=0}^{n-1} a_j 2^j$, it holds that

$$r_{k+1} = \Re(e_{2^n+k}) = \cos\Big(\frac{\pi}{2^{n+1}} + \pi \sum_{j=0}^{n-1} a_j 2^{-j}\Big). \tag{6.5.3}$$

We observe that in such case, we may define the sequence R by $R := (r_k = \cos\phi_k)_{k\geq 0}$ where the sequence of angles $(\phi_k)_{k\geq 0}$ is defined recursively by $\phi_0 = 0$, $\phi_1 = \pi$, $\phi_2 = \frac{\pi}{2}$ and

$$\phi_{2k+1} = \frac{\phi_{k+1}}{2}, \qquad \phi_{2k+2} = \phi_{2k+1} + \pi, \quad k \geq 1. \tag{6.5.4}$$

This recursion provide a very fast and simple process to construct an ℜ-Leja sequence.

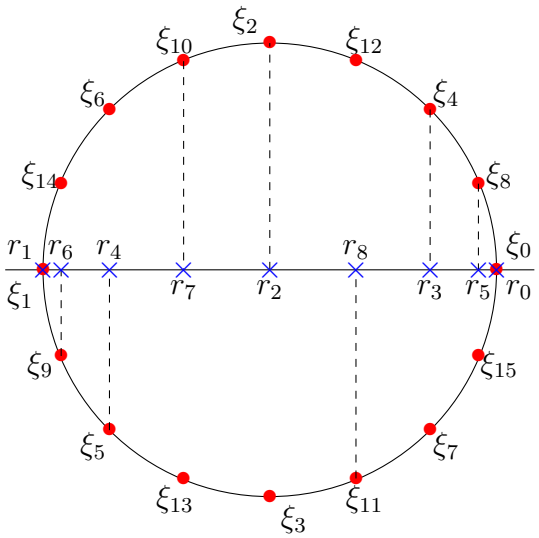

Figure 6.5.2: Distribution of the first 8 elements of the ℜ-Leja sequence R.

6.5.2 Symmetry properties of ℜ-Leja sequences on the unit interval

ℜ-Leja sequences possess interesting symmetry properties that are to some extent inherited from the structural symmetry properties of Leja sequences on the unit disk. First, since a Leja sequence $E=(e_j)_{j\geq 0}$ with initial value $e_0=1$ satisfies $e_1=-1$, $e_2=\pm i$, $e_3=-e_2,\cdots$ and the property $e_{2j+1}=e_{2j}$ for every $j\geq 0$ then the corresponding ℜ-Leja sequence satisfies

$$r_0=1,\ \ r_1=-1,\ \ r_2=0 \quad \text{and} \quad r_{2j}=-r_{2j-1},\ \ j\geq 2. \tag{6.5.5}$$

By the main property of Leja sequences on the unit disk identified in Theorem 6.3.1, given an integer $n\geq 0$, the section $E_{2^{n+1}}$ coincides as a set with $\mathcal{U}_{2^{n+1}}$ the set of 2^{n+1}-roots of unity, therefore, the projection of $E_{2^{n+1}}$ onto $[-1,1]$ yields the set of Gauss Lobatto abscissas of order 2^n, namely

$$R_{2^n+1}=\left\{v_j^n:=\cos\left(\frac{j\pi}{2^n}\right)\ :\ j=0,\ldots,2^n\right\}, \tag{6.5.6}$$

holds in the set sense. The third and most interesting symmetry property of ℜ-Leja sequences is implied from the symmetry property of Leja sequences identified in (6.3.4). It can be stated as follows

Lemma 6.5.2

Let $R:=(r_j)_{j\geq 0}$ be an ℜ-Leja sequence. The sequence $R^2:=(2r_{2j}^2-1)_{j\geq 0}$ is also an ℜ-Leja sequence.

Proof: We consider $E=(e_j)_{j\geq 0}$ to be a Leja sequence associated with R and recall that by Lemma (6.3.4), the sequences $E^2=(e_{2j}^2)_{j\geq 0}$ is also Leja sequence, here starting at 1 since $e_0=1$. We propose to show that R^2 can be obtained by projection of E^2 onto $[-1,1]$, which finishes the proof. The first two elements of R^2 are 1 and -1, so that we only need to show that (6.5.2) holds with R^2 and E^2. For $n\geq 0$ and $2^n\leq k-1<2^{n+1}$, one has $2^{n+1}\leq(2k-1)-1<2^{n+2}$ so that by (6.5.2),

$$r_{2k-1}=\Re(e_{2^{n+1}+2k-1-1})=\Re(e_{2(2^n+k-1)}).$$

Since $2k\geq 4$, then $r_{2k}=-r_{2k-1}$, then

$$2r_{2k}^2-1=2r_{2k-1}^2-1=\Re(e_{2(2^n+k-1)}^2),$$

where we have used $\Re(z^2)=2\Re(z)^2-1$ for $z\in\partial\mathcal{U}$. The proof is then complete. ■

The previous lemma has certain implications on the polynomials w_{R_k} associated with the sections R_k which are very essential on the study of the growth of Lebesgue constants. In order to lighten our notation, we find it convenient to work with normalized versions of the polynomials w_{R_k} that we define by

$$W_{R_k}(x)=2^k w_{R_k}(x),\quad x\in[-1,1]. \tag{6.5.7}$$

In the same fashion of (6.3.6) in the complex case, we are interested in the relation between the polynomial defined in (6.5.7) for sections of the sequences R and R^2. First, since all ℜ-Leja sequences have initial elements 1 and -1, then it is immediate that

$$W_{R_1^2}(2x^2-1) = W_{R_2}(x) \quad x \in [-1,1]. \tag{6.5.8}$$

For higher value of k, we have the following

Lemma 6.5.3

Let R by an ℜ-Leja sequence and $R^2 := \Big(s_j := 2r_{2j}^2 - 1\Big)_{j\geq 0}$. For any $k \geq 2$

$$W_{R_k^2}(2x^2-1) = 2x\, W_{R_{2k-1}}(x), \quad x \in [-1,1] \tag{6.5.9}$$

Consequently $W'_{R_k^2}(-1) = W'_{R_{2k-1}}(0)$, $W'_{R_k^2}(1) = \frac{1}{2}W'_{R_{2k-1}}(1) = \frac{1}{2}W'_{R_{2k-1}}(-1)$ and

$$W'_{R_k^2}(s_j) = \frac{1}{2}W'_{R_{2k-1}}(r_{2j}) = \frac{1}{2}W'_{R_{2k-1}}(r_{2j-1}), \quad j = 2,\ldots,k-1 \tag{6.5.10}$$

Proof: From the definition of R^2 and the property $r_{2j} = -r_{2j-1}$ for $j \geq 2$, we infer that if $k \geq 3$

$$w_{R_k^2}(2x^2-1) = 2^k \prod_{j=0}^{k-1}(x+r_{2j})(x-r_{2j}) = 2^k(x+1)(x-1)x^2\prod_{j=2}^{k-1}(x-r_{2j-1})(x-r_{2j}) = 2^k x\, w_{R_{2k-1}}(x),$$

which implies (6.5.9) for $k \geq 3$. The verification for $k = 2$ is immediate. The derivation with respect to x gives

$$4x\, W'_{R_k^2}(2x^2-1) = 2\Big(x\, W'_{R_{2k-1}}(x) + W_{R_{2k-1}}(x)\Big). \tag{6.5.11}$$

Since $W_{R_{2k-1}}(0) = 0$, then the first result on derivatives is obtained by dividing by x and letting $x \to 0$. The second result is obtained by the substitution of x by 1 or -1. As for (6.5.10), we substitute x by r_{2j} and $r_{2j-1} = -r_{2j}$ for $j = 2\ldots,k-1$. ■

The previous Lemma has in particular an implication on the growth of $W_{R_k}(r_k)$ that we use in the next section

Lemma 6.5.4

Let R be an ℜ-Leja section and denote $S := R^2$. For any $k \geq 2$, if $k = 2N+1$ is odd number, then $2r_k\, W_{R_k}(r_k) = W_{S_{N+1}}(s_{N+1})$ and if $k = 2N$ is an even number, then

$$W_{R_k}(r_k) = 2W_{S_N}(s_N) \tag{6.5.12}$$

Proof: The first equality follows from the previous lemma since $k = 2(N+1)-1$ and therefore $2r_k^2 - 1 = 2r_{2(N+1)}^2 - 1 = s_{N+1}$. As for the second equality, it can be checked easily for $N = 1$ and for $N \geq 2$. Using the fact $r_k = -r_{2N-1}$ and the previous lemma, we infer that

$$W_{R_k}(r_k) = 2(r_k - r_{2N-1})W_{R_{2N-1}}(r_k) = 4r_k W_{R_{2N-1}}(r_k) = 2W_{S_N}(2r_{2N}^2 - 1) = 2W_{S_N}(s_N).$$ ∎

6.5.3 Growth of Lebesgue constant of ℜ-Leja sequences.

The analysis of the growth of Lebesgue constants of ℜ-Leja sequences is quite different from the analysis for Leja sequences on the unit disk. Indeed, ℜ-Leja sequences do not satisfy the Leja definition (6.2.5) on $[-1,1]$, therefore, the machinery developed in complex setting does not apply. However, the "*binary*" structure of Leja sequences on the unit disk convey interesting symmetry properties to ℜ-Leja sequences that we might exploit in order to simplify the analysis. For instance, for the values $k = 2^n + 1$, the section R_{2^n+1} coincides as set with the Gauss Lobatto points, see (6.5.6). This type of abscissas tend to accumulate to the boundaries of $[-1,1]$ as with Tchybeshev abscissas and are known to have optimal Lebesgue constant, in the sense of $\mathbb{L}_k \sim \frac{2}{\pi}\log(k)$, more precisely, we have the bound

$$\mathbb{L}_{R_{2^n+1}} \leq 1 + \frac{2}{\pi}\log(2^n). \tag{6.5.13}$$

See [47, Formulas 5 and 13]. For more general value of k, we propose to relate the analysis of the Lebesgue constant $\mathbb{L}_{R_k}$ to the analysis of the Lebesgue constant $\mathbb{L}_{G_k}$ where G_k is the shortest Leja section on $\mathcal{U}$ that yields R_k when projected onto $[-1,1]$ and take benefit from the machinery developed for the complex setting.

Applying a straightforward cardinality argument, it can be seen from (6.5.1) that for any $n \geq 0$ and any k with $2^n + 1 < k < 2^{n+1} + 1$,

$$R_k = R_{2^n+1} \wedge \Re(E_{2^{n+1},2^{n+1}+k'-1}), \quad k' := k - (2^n + 1). \tag{6.5.14}$$

This shows that for such k, $G_k := E_{2^{n+1}+k'} = E_{k+2^n-1}$ is the section of E of minimal length that yields R_k when projected onto $[-1,1]$. We shall prove the following theorem on the growth of the Lebesgue constants of ℜ-Leja sequences.

Theorem 6.5.5

Let R an ℜ-Leja sequence and E an associated Leja sequence on $\mathcal{U}$. For $n \geq 0$ and $k \geq 3$ such that $2^n + 1 < k < 2^{n+1} + 1$, we have

$$\mathbb{L}_{R_k} \leq 2^{3/2+n-p(k')}\mathbb{L}_{G_k} \quad \text{where} \quad k' = k - (2^n + 1). \tag{6.5.15}$$

where $G_k := E_{2^{n+1}+k'} = E_{k+2^n-1}$.

The previous theorem combined with (6.4.19) implies in particular that

$$\mathbb{L}_{R_k} \leq 2^{5/2}\frac{2^n(k+2^n-1)}{4^{p(k')}}\mathbb{L}_{2^{p(k')}} \leq 8\sqrt{2}k^2 \tag{6.5.16}$$

Clearly, in order to prove the theorem, we should relate the Lebesgue functions associated with the real section R_k and the complex section G_k. In others word, we must investigate how the Lagrange polynomials associated with R_k can be bounded using the Lagrange polynomials associated with G_k. To this end, we explore how the section G_k can be constructed knowing its projection R_k.

Lemma 6.5.6

Consider $E = (e_j)_{j\geq 0}$ a Leja sequence with $e_0 = 1$ and $R := (r_j)_{j\geq 0}$ the associated $\Re$-Leja sequence, and $Z = (z_j)_{j\geq 0}$ the sequence as in 6.5.1. For $n \geq 0$ and k such that $2^n + 1 < k < 2^{n+1} + 1$, we have

$$G_k = \{z_0, z_1, z_2, \overline{z_2}, \cdots, z_{2^n}, \overline{z_{2^n}}\} \cup \{z_{2^n+1}, \cdots, z_{k-1}\}. \tag{6.5.17}$$

in the set sense.

Proof: We have that $G_k = E_{2^{n+1}} \wedge E_{2^{n+1},2^n+k-1}$. Applying cardinality considerations to the definition of Z we infer $E_{2^{n+1},2^n+k-1} = Z_{2^n+1,k}$. Therefore, we only need to show that $E_{2^{n+1}} = \{z_0, z_1, z_2, \overline{z_2}, \cdots, z_{2^n}, \overline{z_{2^n}}\}$. By Theorem 6.3.1, $E_{2^{n+1}}$ coincides with the set of 2^{n+1}- root of unit, therefore E_{2^n+1} is the union of $\{1, -1\}$ and $\{z_2, \ldots, z_{2^n}\}$ and theirs conjugates, which finishes the proof. ■

In view of the above, we can relate the polynomials W_{R_k} and w_{G_k} and theirs derivatives.

Lemma 6.5.7

Let n, k and G_k be as in the previous lemma. For $z \in \partial\mathcal{U}$ and $x = \Re(z)$

$$|W_{R_k}(x)| = |z^2 - 1||w_{G_k}(z)||w_{\overline{F_k}}(z)|. \tag{6.5.18}$$

where $F_k = \{z_{2^n+1}, \ldots, z_{k-1}\}$. Consequently, for $j = 0, \cdots, k-1$

$$|W'_{R_k}(r_j)| = 2\alpha_j|w'_{G_k}(z_j)||w_{\overline{F_k}}(z_j)|, \tag{6.5.19}$$

where $\alpha_j = 1$ for every j except for $j = 0$ and $j = 1$, it is equal to 2.

Proof: Given $x = \frac{1}{2}(z + \overline{z})$ and $x' = \frac{1}{2}(z' + \overline{z}')$, $|z| = |z'| = 1$, one easily check that

$$2|x - x'| = |z - z'||z - \overline{z'}|. \tag{6.5.20}$$

Since $r_j = \Re(z_j)$ and $z_j \in \partial\mathcal{U}$ for any $j \geq 0$, then

$$|W_{R_k}(x)| = \prod_{j=0}^{k-1} 2|x - r_j| = \prod_{j=0}^{k-1} |z - z_j| \prod_{j=0}^{k-1} |z - \overline{z_j}|.$$

In view of $z_0 = 1$, $z_1 = -1$ and the identity (6.5.17), the first result follows. This result combined with the identity (6.5.20), shows that for every $j = 1, \cdots, k-1$

$$|W'_{R_k}(r_j)| = \lim_{x \to r_j} \frac{|W_{R_k}(x)|}{|x - r_j|} = \lim_{z \to z_j} \frac{|z^2 - 1||w_{G_k}(z)||w_{\overline{F_k}}(z)|}{\frac{1}{2}|z - z_j||z - \overline{z_j}|}, \tag{6.5.21}$$

where the limit $\lim_{z \to z_j}$ is meant in the circle $\partial\mathcal{U}$. The second result follows then from the fact that $\lim_{z \to \xi} |z^2 - 1|/|z - \overline{\xi}|$ is equal to 1 for every $\xi \in \partial\mathcal{U}$ expect for $\xi = 1$ and $\xi = -1$, it is equal to 2. ■

In view of the above, we are now able to give the formula giving the Lagrange polynomials associated with the sections R_k and the set G_k and provide the proof of Theorem 6.5.5.

Proof of Theorem 6.5.5 : We denote by $l_0, \cdots, l_{k-1}$ the Lagrange polynomials associated with R_k and by $L_0, L_1, L_{(2,1)}, L_{(2,2)}, \cdots, L_{(2^n,1)}, L_{(2^n,2)}, L_{2^n+1}, \cdots, L_{k-1}$ the Lagrange polynomials associated with the set G_k as given in (6.5.17). We propose to relate the polynomial l_0 to L_0, the polynomial l_1 to L_1, the polynomial l_j to $L_{(j,1)}$ and $L_{(j,2)}$ for every $j = 2, \cdots 2^n$ and the polynomial l_j to L_j for every $j = 2^n + 1, \cdots k-1$. We note that for every $j = 0, \ldots, k-1$, we have

$$l_j(x) = \frac{W_{R_k}(x)}{W'_{R_k}(r_j)(x - r_j)}, \quad x \in [-1, 1]. \tag{6.5.22}$$

Therefore, combining (6.5.18), (6.5.19) and the identity (6.5.20), we deduce that for $j = 0, \ldots, k-1$

$$|l_j(x)| = \frac{1}{\alpha_j} \Big| \frac{z^2 - 1}{(z - z_j)(z - \overline{z_j})} \Big| \frac{|w_{G_k}(z)||w_{\overline{F_k}}(z)|}{|w'_{G_k}(z_j)||w_{\overline{F_k}}(z_j)|}, \quad z \in \partial\mathcal{U}, x = \Re(z). \tag{6.5.23}$$

where α_j are defined as in 6.5.7. To lighten our notations, we introduce the quotients

$$q_k(z, \xi) := \frac{|w_{\overline{F_k}}(z)|}{|w_{\overline{F_k}}(\xi)|}, \qquad z \in \partial\mathcal{U}, \xi \in \partial\mathcal{U} \setminus F_k. \tag{6.5.24}$$

Using elementary rational decomposition, with variable ξ and constant $z \in \partial\mathcal{U}$, it can be shown that

$$\Big| \frac{z^2 - 1}{(z - \xi)(z - \overline{\xi})} \Big| = \Big| \frac{z - \overline{z}}{(z - \xi)(\overline{z} - \xi)} \Big| \leq \frac{1}{|z - \xi|} + \frac{1}{|z - \overline{\xi}|} = \frac{1}{|z - \xi|} + \frac{1}{|\overline{z} - \xi|} \tag{6.5.25}$$

This last inequality applied with the real values $\xi = z_0 = 1$ and $\xi = z_1 = -1$ and injected in (6.5.23) yields

$$|l_0(x)| \leq q_k(z, z_0)|L_0(z)| \quad \text{and} \quad |l_1(x)| \leq q_k(z, z_1)|L_1(z)| \tag{6.5.26}$$

For the indices $j = 2, \ldots, 2^n$, since z_j and $\overline{z_j}$ play symmetric role in that $\Re(z_j) = \Re(\overline{z_j}) = r_j$ and $z_j, \overline{z_j} \in G_k$, then one observes that (6.5.19) yields $|w'_{G_k}(z_j)||w_{\overline{F_k}}(z_j)| = |w'_{G_k}(\overline{z_j})||w_{\overline{F_k}}(\overline{z_j})|$. Therefore, taking account of this equality when injecting the first part of (6.5.25) in (6.5.23), we deduce

$$|l_j(x)| \leq q_k(z, z_j)L_{(j,1)}(z) + q_k(z, \overline{z_j})L_{(j,2)}(z), \tag{6.5.27}$$

Finally for the indices $j = 2^n + 1, \ldots, k-1$, we inject the second inequality of (6.5.25) in (6.5.23) but taking account of $|w_{G_k}(z)w_{\overline{F_k}}(z)| = |w_{G_k}(\overline{z})||w_{\overline{F_k}}(\overline{z})|$, which follows from $G_k \cup \overline{F_k} = \overline{G_k} \cup F_k$, we obtain

$$|l_j(x)| \leq q_k(z, z_j)L_j(z) + q_k(\overline{z}, z_j)L_j(\overline{z}). \tag{6.5.28}$$

Combing the inequalities (6.5.26), (6.5.27) and (6.5.28), we infer the rough bound

$$\mathbb{L}_{R_k} \leq 2\mathbb{L}_{G_k} \sup_{\substack{z \in \partial\mathcal{U} \\ \xi \in G_k}} q_k(z, \xi). \tag{6.5.29}$$

By the structure of Leja sequences on $\mathcal{U}$, we have that $F_k = E_{2^{n+1}, 2^n+k-1}$ is a k'-Leja section with $k' = k - (2^n + 1)$ and $0 < k' < 2^n$, therefore by Corollary 6.3.7, we deduce

$$q_k(z, \xi) = \frac{|w_{F_k}(\overline{z})|}{|w_{F_k}(\overline{\xi})|} \leq \frac{2^{\sigma_1(k')}2^{\sigma_1(2^n-k')}}{\overline{\xi}^{2^n} - e_{2^{n+1}}^{2^n}} = \frac{2^{n+1-p(k')}}{\overline{\xi}^{2^n} - e_{2^{n+1}}^{2^n}}$$

Since $e_{2^{n+1}}$ is a 2^{n+1}-root of -1, then $e_{2^{n+1}}^{2^n} = \pm i$. As for $\xi \in G_k$, since $G_k \subset E_{2^{n+2}} = \mathcal{U}_{E_{2^{n+2}}}$ then $\xi^{2^n} \in \{1, -1, i, -i\}$. This shows that necessarily $|\overline{\xi}^{2^n} - e_{2^{n+1}}^{2^n}| \geq \sqrt{2}$, so that

$$\sup_{\substack{z \in \partial\mathcal{U} \\ \xi \in G_k}} q_k(z, \xi) \leq 2^{n+\frac{1}{2}-p(k')} \tag{6.5.30}$$

This bound injected in (6.5.29) completes the proof of Theorem 6.5.5. ■

Remark 6.5.8

The previous approach can be applied to bound the Lebsgue constant associated with the Gauss-Lobatto abscissas $T_k^* := \{\cos\frac{j\pi}{k} : j = 0, \cdots, k\}$ *using the Lebesgue constant associated with the* $2k$*-roots of unity* $\mathcal{U}_{2k} := \{\exp(\frac{j}{k}i\pi) : j = 0, \cdots, 2k-1\}$. *Indeed, the same arguments shows that*

$$\mathbb{L}_{T_k^*} \leq \mathbb{L}_{\mathcal{U}_{2k}} \tag{6.5.31}$$

Using the known result $\mathbb{L}_{\mathcal{U}_N} \leq \frac{2}{\pi}\log N + \mathcal{O}(1)$, *this shows that* $\mathbb{L}_{T_k^*} \leq \frac{2}{\pi}\log k + \mathcal{O}(1)$ *which is a new approach to prove this classical result.*

6.6 Norms of the difference operators

In this section, we focus our attention on the difference operators associated with interpolation on nested sequences. We first define theses operators on an abstract setting, then we analyze them in the case of Leja sequences on $\mathcal{U}$ and $\Re$-Leja sequences on $[-1,1]$.

We consider $Z=(z_j)_{j\geq 0}$ a sequence of pairwise distinct points in a compact set X contained either in $\mathbb{R}$ or $\mathbb{C}$ and denote by Π_{Z_k} the interpolation operators associated with the sections Z_k. We introduce the difference operators $(\Delta_j)_{j\geq 0}$ defined by

$$\Delta_0=\Pi_{Z_1},\quad \text{and}\quad \Delta_k=\Pi_{Z_{k+1}}-\Pi_{Z_k},\quad k\geq 1. \tag{6.6.1}$$

We are interested in the norm of theses operator defined by

$$\delta_k:=\sup_{g\in C(X)-\{0\}}\frac{\|\Delta_k g\|_{L^\infty(X)}}{\|g\|_{L^\infty(X)}} \tag{6.6.2}$$

We may write $\delta_k(Z)$ to emphasize the dependence on the sequence Z. It is immediate that $\delta_0=\mathbb{L}_{Z_1}=1$ and $\delta_k\leq \mathbb{L}_{Z_{k+1}}+\mathbb{L}_{Z_k}$ for $k\geq 1$. We shall sharpen the previous bound when Z has a particular structure, for instance, if Z is a Leja or an $\Re$-Leja sequence. As for Lebesgue constant, we can express δ_k using Lagrange polynomials. Indeed, using Lagrange interpolation formula with the section Z_{k+1}, it can be easily checked that for any $k\geq 1$

$$\Delta_k g(z)=\Big(g(z_k)-\Pi_{Z_k}g(z_k)\Big)\frac{w_{Z_k}(z)}{w_{Z_k}(z_k)},\quad z\in X. \tag{6.6.3}$$

This implies that

$$\delta_k=\sup_{z\in X}\frac{|w_{Z_k}(z)|}{|w_{Z_k}(z_k)|}\sup_{f\in C(X)-\{0\}}\frac{|g(z_k)-\Pi_{Z_k}g(z_k)|}{\|g\|_{L^\infty(X)}} \tag{6.6.4}$$

The second supremum in the previous equality is obviously bounded by $1+\lambda_{E_k}(z_k)$. This bound is actually attained, to see this consider g to be a function in $C(X)$ having a maximum value equal to 1, and satisfying $g(z_k)=-1$ and $g(z_j)=\frac{|l_j(z_k)|}{l_j(z_k)}$ for every $j=0,\dots,k-1$ where $l_0,\dots,l_{k-1}$ are the Lagrange polynomials associated with E_k. Therefore

$$\delta_k=\Big(1+\lambda_{E_k}(z_k)\Big)\sup_{z\in X}\frac{|w_{Z_k}(z)|}{|w_{Z_k}(z_k)|}. \tag{6.6.5}$$

The previous formula shows in particular

$$Z \text{ is a Leja sequence on } X \Longrightarrow \delta_k=1+\lambda_{Z_k}(z_k). \tag{6.6.6}$$

In particular, in view of the results on Leja sequences on the unit disk, Lemma 6.4.3, we have the following

Lemma 6.6.1

Let E be a Leja section in $\mathcal{U}$ with initial value $e_0 \in \partial\mathcal{U}$. The difference operators associated with E satisfies, $\delta_0(E) = 1$ and for $k \geq 1$

$$\delta_k(E) \leq 1 + k \tag{6.6.7}$$

The formula (6.6.5) is convenient in the case of Leja sequences since it yields exact values of the quantities δ_k. For the sake of convenience in the case of ℜ-Leja sequences, we opt for a different rearrangement of (6.6.5). From the formulas (6.2.2) of Lagrange polynomials associated with Z_k, we may write (6.6.5) as

$$\delta_k = \Big(\frac{1}{|w_{Z_k}(z_k)|} + \sum_{j=0}^{k-1} \frac{1}{|w'_{Z_k}(z_j)||z_k - z_j|}\Big) \sup_{z\in X} |w_{Z_k}(z)|. \tag{6.6.8}$$

We remark that $|w_{Z_k}(z_k)| = |w'_{Z_{k+1}}(z_k)|$ and $|w'_{Z_k}(z_j)||z_k - z_j| = |w'_{Z_{k+1}}(z_j)|$ for any $j = 0, \ldots, k-1$, we may then rewrite (6.6.5) in the more compact form

$$\delta_k = \Big(\sum_{j=0}^{k} \frac{1}{|w'_{E_{k+1}}(z_j)|}\Big) \sup_{z\in X} |w_{Z_k}(z)| \tag{6.6.9}$$

Now giving $R = (r_j)_{j\geq 0}$ an ℜ-Leja sequence on $[-1, 1]$, using the polynomials $W_{R_k} = 2^k w_{R_k}$ defined in (6.5.7), the previous formula becomes

$$\delta_k(R) = 2\beta_k(R) \sup_{x\in[-1,1]} |W_{R_k}(x)| \qquad \text{where} \qquad \beta_k(R) := \sum_{j=0}^{k} \frac{1}{|W'_{R_{k+1}}(r_j)|}. \tag{6.6.10}$$

We propose to bound the quantity $\beta_k(R)$ for any ℜ-Leja sequence R.

Lemma 6.6.2

Let R by an ℜ-Leja sequence. We have $\beta_{2^n}(R) = \frac{1}{4}$ for any $n \geq 0$ and for $k \geq 1$, such that $2^n < k < 2^{n+1}$,

$$\beta_k(R) \leq C\frac{2^{\sigma_0(k)}}{2^{p(k)}}, \qquad C = \frac{1}{2}. \tag{6.6.11}$$

where $\sigma_0(k)$ is the number of zeros in the binary expansion of k.

Proof: First, we assume that $k = 2N \geq 4$ is an even integer. We have

$$\beta_k(R) = \frac{1}{|W'_{R_{2N+1}}(1)|} + \frac{1}{|W'_{R_{2N+1}}(-1)|} + \frac{1}{|W'_{R_{2N+1}}(0)|} + \sum_{j=2}^{N} \Big(\frac{1}{|W'_{R_{2N+1}}(r_{2j-1})|} + \frac{1}{|W'_{R_{2N+1}}(r_{2j})|}\Big). \tag{6.6.12}$$

We introduce the shorthand $S = R^2$. Using Lemma 6.5.3, we deduce that

$$\beta_k(R) = \frac{1}{|W'_{S_{N+1}}(1)|} + \frac{1}{|W'_{S_{N+1}}(-1)|} + \sum_{j=2}^{N} \frac{1}{|W'_{S_{N+1}}(s_j)|} = \beta_N(S). \tag{6.6.13}$$

The same arguments implies that $\beta_2(R) = \beta_1(S)$, so that $\beta_{2N}(R) = \beta_N(S)$ is valid for any $N \geq 1$. Since S is also an $\Re$-Leja sequence, then the verification $\beta_1(S) = \frac{1}{4}$ for any $\Re$-Leja sequence S implies the first result in the lemma $\beta_{2^n}(R) = \frac{1}{4}$ for any $n \geq 0$.

To prove the second part of the lemma, we use induction on $k \geq 3$. First, since $r_0 = 1$, $r_1 = -1$, $r_2 = 0$ and $r_3 = \pm\sqrt{2}$, then it is easily checked that $\beta_3(R) = \frac{\sqrt{2}}{8}$, satisfies (6.6.11). Let now $k \geq 3$ and assume the induction hypothesis holds for any $j < k$. If $k = 2N$ is even with N not a power of 2, then $N \geq 3$, therefore (6.6.13) combined with the induction hypothesis with N implies the result for k since $\sigma_0(N) = \sigma_0(k) - 1$ and $p(N) = p(k) - 1$.

We now assume that $k = 2N + 1 \geq 5$ is an odd integer. First, we isolate the last quotient in the the sum giving $\beta_k(R)$ and multiply the other quotients by $\frac{|r_j - r_{k+1}|}{|r_j - r_{k+1}|}$ yielding

$$\beta_k(R) = \frac{1}{W_{R_k}(r_k)} + \sum_{j=0}^{k-1} \frac{|r_j - r_{k+1}|}{|W'_{R_{k+2}}(r_j)|}.$$

Since $k = 2(N+1)-1$ and $k+2 = 2(N+2)-1$, then regrouping the sum as in (6.6.12) and using Lemma 6.5.3, taking into account $r_0 = 1$, $r_1 = -1$ and $r_2 = 0$, we deduce

$$\begin{aligned}\beta_k(R) = \frac{2|r_k|}{|W_{S_{N+1}}(s_{N+1})|} + & \frac{|1 - r_{2N+2}| + |-1 - r_{2N+2}|}{2|W'_{S_{N+2}}(1)|} + \frac{|r_{2N+2}|}{|W'_{S_{N+2}}(-1)|} + \\ & \Big(\sum_{j=2}^{N} \frac{|r_{2j-1} - r_{2N+2}| + |r_{2j} - r_{2N+2}|}{2|W'_{S_{N+2}}(s_j)|}\Big)\end{aligned}$$

Since $|x - r| + |x + r| \leq 2$ for any $x, r \in [-1, 1]$ and $r_{2j-1} = -r_{2j}$, for every $j \geq 2$, we deduce that

$$\begin{aligned}\beta_k(R) \leq \frac{2}{|W_{S_{N+1}}(s_{N+1})|} + & \frac{1}{|W'_{S_{N+2}}(1)|} + \frac{1}{|W'_{S_{N+2}}(-1)|} + \sum_{j=2}^{N} \frac{1}{|W'_{S_{N+2}}(s_j)|} \\ & = \frac{1}{|W_{S_{N+1}}(s_{N+1})|} + \beta_{N+1}(S) \leq 2\beta_{N+1}(S)\end{aligned}$$

First, if $k + 1$ is a power of 2, i.e. $k = 2^q - 1$, $q \geq 1$, then $N + 1 = 2^{q-1}$, so that the previous inequality implies $\beta_k(R) \leq \frac{1}{2}$ which is compatible with (6.6.11) since $\sigma_0(k) = p(k) = 0$. Now if $k + 1$ is not a power of 2, we write $k = \sum_{j=0}^{n} a_j 2^j$ and introduce $q = \inf\{j \geq 0 : a_j = 0\}$. We have $k = 2^q - 1 + 2^{q+1}m$ with $m \geq 1$. The induction hypothesis applied with $N+1 = 2^{q-1}(1+2m)$ combined with (6.6.13) yields

$$\beta_k(R) \leq 2\beta_{2^{q-1}(1+2m)}(R^2) = 2\beta_{1+2m}(R^{2^q}) \leq 2C\frac{2^{\sigma_0(1+2m)}}{2^{p(1+2m)}} = C2^{\sigma_0(k)},$$

where we have used that $\sigma_0(1+2m) = \sigma_0(m) = \sigma_0(k) - 1$ since $k = \underbrace{11\ldots1}_{q}0m$ in the sense of binary expansions and that R^{2q} is an ℜ-Leja sequence. The proof is now complete. ■

In view of the above lemma, we are now able to provide a bound on the growth of the norms of the difference operators for ℜ-Leja sequence.

Lemma 6.6.3

Let R be an ℜ-Leja sequences in $[-1,1]$. For any $n \geq 0$ and for $k \geq 1$, such that $2^n \leq k < 2^{n+1}$,

$$\delta_k(R) \leq 4^n \tag{6.6.14}$$

Proof: We have by Lemma (6.5.18) that for $2^n + 1 < k < 2^{n+1} + 1$

$$|W_{R_k}(x)| = |z^2-1||w_{G_k}(z)||w_{\overline{F_k}}(z)| \leq 22^{\sigma_1(2^{n+1}+k')}2^{\sigma_1(k')} = 44^{\sigma_1(k')}, \qquad k' = k-(2^n+1)$$

This result is also valid for any k. First, we treat the case $k = 2^n$. We have in such case $2^{n-1} + 1 < k < 2^n + 1$, so that the previous inequality implies $|W_{R_k}(x)| \leq 44^{\sigma_1(2^{n-1}-1)} = 44^{n-1} = 4^n$. This combined with the previous Lemma implies

$$\delta_{2^n}(R) \leq 2\frac{1}{4}4^n \leq 4^n$$

For k not a power of 2, we have since $0 < k' < 2^n$, the number of ones in the binary expansion of k' satisfies $\sigma_1(k') = \sigma(k'+2^n) - 1 = \sigma(k-1) - 1$. It can be checked using binary subtraction $\sigma_1(k-1) = \sigma_1(k) - 1$ if k is odd and $\sigma_1(k-1) = p(k) - 1 + \sigma_1(k)$ for k even, therefore

$$\sigma_1(k') = \sigma_1(k) + p(k) - 1$$

We deduce then from (6.6.10) and the previous lemma that

$$\delta_k(R) \leq 4^{\sigma_1(k)+p(k)-1}2^{\sigma_0(k)}2^{-p(k)} = \frac{1}{4}2^{\sigma_1(k)+p(k)}2^{\sigma_1(k)+\sigma_0(k)} \leq \frac{1}{4}(2^{n+1})^2 = 4^n.$$

where we have used $\sigma_1(k) + p(k) \leq \sigma_1(k) + \sigma_0(k) = n + 1$. ■

6.7 Numerical illustration

We have computed numerically the Lebesgue constants $\mathbb{L}_k$ of the Leja sections on the unit disk and the Lebesgue constants $\mathbb{L}_{R_k}$ with R is the ℜ-Leja sequence given by (6.5.3), up to the value $k = 129$. Figure 6.7.3 display their behaviours with respect to k.

In the complex case, we notice the regular patterns in the graph of $k \mapsto \mathbb{L}_k$, which reveal the particular role of divisibility by powers of 2 in k. This role also appears in the estimate (6.4.5), due to the presence of 2^{p_k} in the denominator. The worst values of $\mathbb{L}_k$ appear for the values $k = 2^n - 1$ for which it was proved in [18] that $\mathbb{L}_k = k$. The conjecture $\mathbb{L}_k \leq k$ seems reasonable in view of this graph.

In the real case, the patterns are also present yet less visible. One can also see that $\mathbb{L}_{R_k} \geq k$ for certain values of k. However, the graph does not give any clear intuition on the best asymptotic estimate that should be expected.

We may think of other sequences of points in $[-1, 1]$ for which the Lebesgue constant could behave better than for the sequence R. As an example, we have numerically computed, for $k = 1, \ldots, 129$, the Lebesgue constants when using the k-sections of the two following sequences:

- The standard Leja sequence L on $[-1, 1]$ with starting point $r_0 = 1$, which is iteratively built by taking

$$r_j \in \mathrm{Argmax}_{x\in[-1,1]} \prod_{l=0}^{j-1} |x - r_l|. \tag{6.7.1}$$

  The computation of this sequence becomes intensive for larger values of k.

- The sequence M on $[-1, 1]$ with starting point $r_0 = 1$, which is iteratively built by maximization of the Lebesgue function of its sections according to

$$r_j \in \mathrm{Argmax}_{x\in[-1,1]} \lambda_{M_{j-1}}(x), \tag{6.7.2}$$

  where M_j denotes the j-section of M. The computation of this sequence is as intensive as that of L.

Figure 6.7.4 displays the comparison between the Lebesgue constants for the three sequences R, L and M. We observe that the behaviour for the sequences L and M is very similar and generally better than for the sequence R, to the exception of isolated values such as $k = 2^n + 1$ for which the R_k coincide with the Gauss-Lobatto points giving therefore $\mathbb{L}_{R_k} \lesssim \log k$. Note however that we do not hove bounds for $\mathbb{L}_{L_k}$ and $\mathbb{L}_{M_k}$ that are comparable to the quadratic growing bounds obtained for $\mathbb{L}_{R_k}$.

6.8 Conclusion

In this chapter, we have proved that the growths of the Lebesgue constants associated with Leja sequences on unit circle and their projections the $\Re$-Leja sequences are

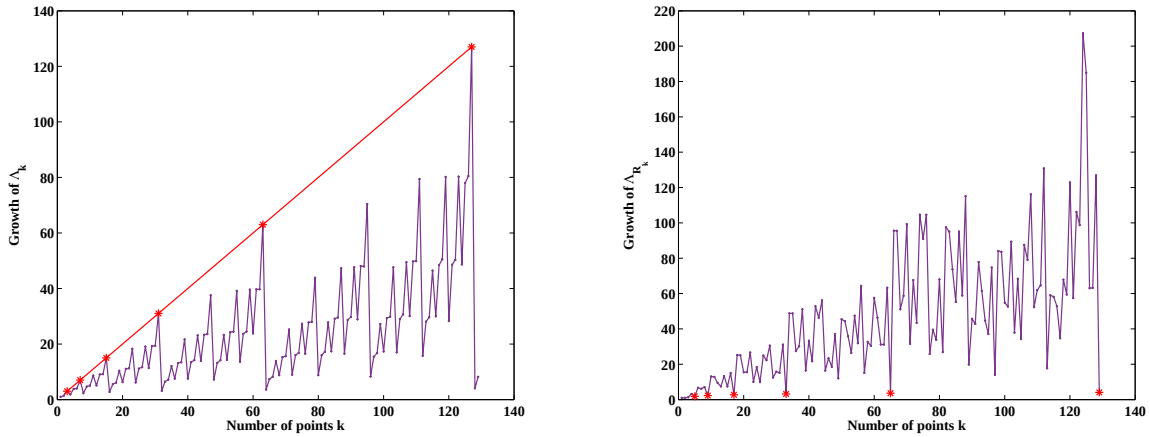

Figure 6.7.3: Exact Lebesgue constants $\mathbb{L}_k$ (left) and $\mathbb{L}_{R_k}$ (right) for $k = 1, \ldots, 129$.

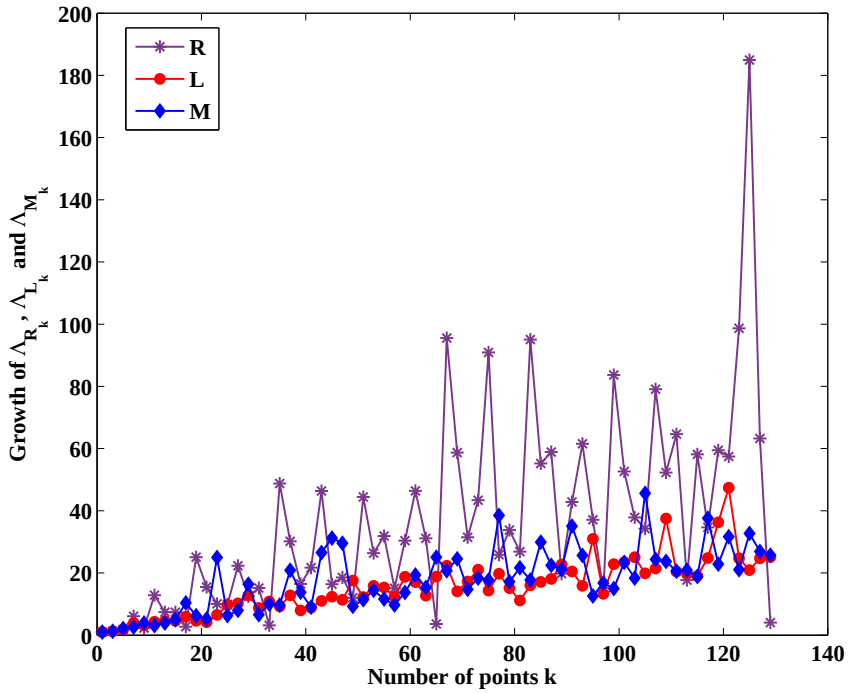

Figure 6.7.4: Exact Lebesgue constants associated to the k-sections of R, L and M, for $k = 1, 3, \ldots, 129$.

respectively sub-linear and sub-quadratic. We have shown in addition that the associated difference operator Δ_k have norms which are bounded in $k + 1$ and $(k + 1)^2$

respectively. This implies in view of Remark 5.3.2 that the multi-dimensional polynomial interpolation operators, introduced in Chapter 5, based on such sequences have Lebesgue constant that are bounded in $(\#(\Lambda))^2$ and $(\#(\Lambda))^3$ respectively, see (5.2.6) in Chapter 5.

For the purpose of our analysis in this chapter, we have exhibited many interesting structural properties of both the Leja and $\Re$-Leja sequences. In addition, we have introduced a new approach for the study of Lebesgue constants of $\Re$-Leja sequences. The linear bound $2k$ for Leja sequences is somewhat satisfactory since the value k can be attained for integer of the form $2^n - 1$, however for $\Re$-Leja sequence, the bound $8k^2$ is in some way pessimistic. It is of interest to investigate if linear bounds hold also in this case.

Chapter 7

Sparse high-dimensional polynomial least-squares

Contents

7.1 Introduction

In this chapter, we introduce a simple least-squares scheme which can be used for non-intrusive treatment of parametric PDEs. We study a standard polynomial least-squares process which turns out to be stable for projection spaces $\mathbb{V}_\Lambda$ for Λ lower.

As in Chapter 5, we are interested in parametric PDEs of the general form

$$\mathcal{D}(u, y) = 0, \tag{7.1.1}$$

where $u \mapsto \mathcal{D}(u, y)$ is a partial differential operator that depends on a parameter vector $y = (y_j)_{j\geq 1}$ which varies in the parametric domain $U = [-1, 1]^{\mathbb{N}}$. We assume that the

problem (7.1.1) is well posed in some Banach space V for any $y \in U$, so that we may defined the solution map u by

$$y \in U \mapsto u(y) \in V. \tag{7.1.2}$$

We assume that $\mathcal{D}$ satisfies a (p, ε)-holomorphy assumption, see Definition 2.2.1, for some $0 < p < 1$. As a consequence, there exist sequences of nested lower sets $(\Lambda_n^P)_{n\geq 1}$ with $\#(\Lambda_n) = n$ such that u can be approximated in the spaces $\mathbb{V}_{\Lambda_n^P}$ in the uniform sense with algebraic rates,

$$\inf_{v\in\mathbb{V}_{\Lambda_n^P}} \|u - v\|_{\mathcal{V}_\infty} \leq C(n+1)^{-s}, \qquad s := \frac{1}{p} - 1, \tag{7.1.3}$$

or up to considering a different sequence $(\Lambda_n^L)_{n\geq 1}$, also in the mean squares sense with a better rate,

$$\inf_{v\in\mathbb{V}_{\Lambda_n^L}} \|u - v\|_{\mathcal{V}_2} \leq C'(n+1)^{-s^*}, \qquad s^* := \frac{1}{p} - \frac{1}{2}, \tag{7.1.4}$$

As discussed in Chapter 5, we have seen multiple polynomial approximation methods. For instance, Taylor series, Galerkin projection, interpolation and sparse grids collocation method. Such methods provide computable approximation of u in the sense of (7.1.3). However, computable approximation of u in the sense of (7.1.4) are only available for the elliptic model, based on Galerkin projection, see Chapter 4. A typical challenge is then to compute approximation in the sense (7.1.4) for more general models.

The objective of this chapter is to propose and study a collocation method based on a high dimensional least-squares process using observation of the solution map u on independent and identically distributed copies of the random vector y. Throughout this chapter, we only work with lower sets Λ, which once again we recall are defined according to

$$\nu \in \Lambda \quad \text{and} \quad \mu \leq \nu \Rightarrow \mu \in \Lambda, \tag{7.1.5}$$

The convergence analysis in least-squares sense of collocation methods is less satisfactory in the sense that convergence rates similar to (7.1.4) do not seem to have been established for such methods. This is in part due to the difficulty to control the least-squares projection for general multivariate polynomial spaces. We have seen in Chapter 5 that the convergence rate in (7.1.3) can be achieved if interpolation is used with carefully selected points. least-squares methods have been recently analyzed in [32, 66] in the stochastic setting, assuming that the samples y^i are independent realizations of the random variable y, therefore identically distributed according to ϱ. The analysis reveals that in the univariate case where $y \in [-1, 1]$ and for the uniform distribution, the least-squares method is stable and produces a near best approximation in $L^2([-1, 1], \frac{dt}{2})$, under the condition that the number of samples n scales quadratically

(up to a logarithmic factor) with respect to the dimension m of the polynomial space $\mathbb{P}_{m-1}$.

The objective of this chapter is to address the problem of the stability and convergence of the polynomial least-squares method in the general context of the spaces $\mathbb{V}_\Lambda$ associated to arbitrary lower sets Λ. We begin in Section 7.2 by discussing the least squares method for a real-valued function in a general framework not limited to polynomials and recalling recent stability and approximation results established in [32]. In Section 7.3, we focus on the particular framework of the multivariate polynomial spaces $\mathbb{P}_\Lambda$. Our analysis reveals in particular that, with $U = [-1,1]^d$ and the uniform distribution, the same scaling $n \sim \#(\Lambda)^2$ up to a logarithmic factor as in the univariate case, ensures stability and near best approximation of the method *independently of the dimension* d.

In Section 7.4, we show how a similar analysis applies to V-valued functions, where V is a Hilbert space, and therefore to the solutions of parametric and stochastic PDEs. As a relevant example, the equation (1.1.1) with random inclusions in the diffusion coefficient is discussed in §5, and numerical illustration for this example are given in §6.

7.2 Discrete least-squares approximations

Let (U, Θ, ϱ) be a probability space. Here, the domain U and the measure ϱ are not necessarily $[-1,1]^{\mathbb{N}}$ and the uniform measure over $[-1,1]^{\mathbb{N}}$. We denote by $L^2(U, d\varrho)$ the Hilbert space of real-valued squares integrable functions with respect to ϱ and denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the associated inner product and norm, i.e.

$$\langle v, w \rangle := \int_U v(y) w(y) d\varrho(y), \quad \|v\| := \sqrt{\langle v, v \rangle}, \qquad v, w \in L^2(U, \varrho). \tag{7.2.1}$$

We consider X_m a finite dimensional space of $L^2(U, \varrho)$ with $\dim(X_m) = m$. We assume that the functions belonging to X_m are defined everywhere over U. We let $\mathcal{B}_L := (L_j)_{1 \le j \le m}$ be any orthonormal basis of X_m with respect to the above inner product. The best approximation of a function $u \in L^2(U, d\varrho)$ in X_m in the least-squares sense is given by

$$P_m u := \operatorname*{argmin}_{v \in X_m} \int_U |u(y) - v(y)|^2 d\varrho = \sum_{j=1}^m \langle u, L_j \rangle \, L_j, \tag{7.2.2}$$

and its best approximation error by

$$e_m(u) := \inf_{v \in X_m} \|u - v\| = \|u - P_m u\|. \tag{7.2.3}$$

The approximation $P_m u$ is in general out of reach. It requires the knowledge of all the coefficient $\langle u, L_j \rangle$ which in general can only be approximated using quadratures

methods. This however might not produce a stable approximation. Also, in general u is an unknown function, for instance a solution of a given partial differential equation, for which one only has noisy observations. A popular way to approximate u is by virtue of the least-squares method.

If the function u is an unknown function and $(z^i)_{i=1,\cdots,n} \in \mathbb{R}^n$ are noiseless or noisy observations of u at the points $(y^i)_{i=1,\cdots,n} \in U$ where the y^i are i.i.d. random variables distributed according to ϱ. We introduce the discrete least-squares approximation

$$w := \operatorname*{argmin}_{v \in X_m} \sum_{i=1}^{n} |z^j - v(y^j)|^2. \tag{7.2.4}$$

More precisely, the observation model is

$$z^i = u(y^i) + \eta^i, \quad i = 1, \ldots, m, \tag{7.2.5}$$

where y^i are i.i.d. random variable distributed according to ϱ and where η^i represents the noise. Several scenarii may be considered for modeling the noise:

(i) Noiseless model: one has $\eta^i = 0$.

(ii) Stochastic noise model: η^i are centered i.i.d. random variables, with uniformly bounded variance

$$\sup_{y \in \Gamma} \mathbb{E}(|\eta|^2 | y) < \infty. \tag{7.2.6}$$

(iii) Deterministic noise model: $\eta^i = \eta(y_i)$ where η is a uniformly bounded function on Γ with

$$\|\eta\|_{L^\infty(\Gamma)} < \infty \tag{7.2.7}$$

In the framework of parametric PDE's, the observation noise represents the discretization error between the exact solution $u(y)$ and the solution computed by deterministic numerical solver, which is a function of y. The deterministic noise model is therefore the appropriate one, with $\|\eta\|_{L^\infty(\Gamma)}$ representing a uniform bound on the discretization error guaranteed by the numerical solver.

The minimization problem (7.2.4) always has a solution, which may not be unique. In particular, it is never unique in the regime $m > n$. In the following, we only consider the regime $m \leq n$. In the noiseless case (i) above where $z^i = u(y^i)$, the solution may be viewed as the orthogonal projection of u onto X_m with respect to the inner product $\langle \cdot, \cdot \rangle_n$ associated with the empirical semi-norm

$$\|v\|_n = \Big(\frac{1}{n} \sum_{i=1}^{n} |v(y^i)|^2 \Big)^{\frac{1}{2}}. \tag{7.2.8}$$

In this case, we denote the solution w of the problem (7.2.9) by $P_m^n u$, i.e.

$$P_m^n u := \operatorname*{argmin}_{v \in X_m} \sum_{i=1}^{n} |u(y^j) - v(y^j)|^2 = \operatorname*{argmin}_{v \in X_m} \|u - v\|_n. \tag{7.2.9}$$

The projection $P_m^n u$ depends on the sample $(y^j)_{1 \le j \le n}$, so that $P_m^n u$ is a "random" least-squares projector. In both the noisy and noiseless case, the coordinate vector $\mathbf{w} \in \mathbb{R}^m$ of w in the basis $\mathcal{B}_L$ is the solution to the system

$$\mathbf{G}\mathbf{w} = \mathbf{J}\mathbf{z}, \tag{7.2.10}$$

where $\mathbf{G}$, $\mathbf{J}$ and z are respectively the $m \times m$ matrix, the $m \times n$ matrix and the n-dimensional vector of observations all given by

$$\mathbf{G}_{ij} := \langle L_i, L_j \rangle_n, \qquad \mathbf{J}_{ij} := \frac{L_i(y^j)}{n}, \quad \text{and} \quad \mathbf{z}_i = z^i. \tag{7.2.11}$$

Note that $n\mathbf{J}\mathbf{J}^t = G$. In the case where the matrix $\mathbf{G}$ is not singular, the solution w of the least-squares problem (7.2.9) is given by

$$w = \sum_{j=1}^{n} z^j \pi_j. \tag{7.2.12}$$

where $\mathcal{B}_\pi := \{\pi_1, \dots, \pi_n\}$ is a family of elements of X_m given by

$$\mathcal{B}_\pi := \left(\mathbf{G}^{-1}\mathbf{J}\right)^t \mathcal{B}_L, \tag{7.2.13}$$

with the product matrix-basis to be understood in the obvious sense. In the case where $\mathbf{G}$ is singular, we set by convention $w := 0$. If u satisfies a uniform bound $|u(y)| \le L$ over U, where L is known, we introduce the truncated least squares approximation

$$\tilde{w} = T_L(w), \quad \text{with} \quad T_L(t) := \operatorname{sign}(t) \min\{L, |t|\}, \tag{7.2.14}$$

which we also denote by $\tilde{P}_m^n u$ in the noiseless case.

The analysis in [32, 66] investigates the minimal amount of sampling $n = n(m) \ge m$ that allows an accurate approximation of the unknown function u by the random approximations w or $\tilde{w}$. The accuracy here is to be understood in the sense of a comparison between the error $\|u - w\|$ and the best approximation error $e_m(u)$. This analysis is based on probabilistic estimates comparing the norm $\|\cdot\|$ and its empirical counterpart $\|\cdot\|_n$ uniformly over the space X_m. This comparison amounts in estimating the deviation of the random matrix $\mathbf{G}$ from its expectation $\mathbb{E}(\mathbf{G}) = \mathbf{I}$, where $\mathbf{I}$ is the $m \times m$ identity matrix. Indeed, for $v \in X_m$ and $\mathbf{v}$ the vector representing v in the basis $\mathcal{B}_L$, one has

$$\|v\|_n^2 = \mathbf{v}^T \mathbf{G} \mathbf{v} \quad \text{and} \quad \|v\|^2 = \mathbf{v}^T \mathbf{v} = \mathbf{v}^T \mathbf{I} \mathbf{v}, \tag{7.2.15}$$

so that for any $0 < \delta < 1$,

$$|||\mathbf{G} - \mathbf{I}||| \leq \delta \Leftrightarrow |\|v\|_n^2 - \|v\|^2| \leq \delta \|v\|^2, \quad v \in X_m, \tag{7.2.16}$$

where $|||\cdot|||$ denote the spectral norm of a matrix.

We recall in a nutshell the analysis in [32, 66]. First, we define a quantity that plays a central role in both works. Namely, we consider

$$K(X_m) = \sup_{v \in X_m - \{0\}} \frac{\|v\|_{L^\infty(U)}^2}{\|v\|^2}, \tag{7.2.17}$$

The quantity can be viewed as a measure of the stability of the uniform norm with respect to the mean square norm over X_m. We remark that $K(X_m)$ is always greater than 1. By writing $v = \sum_{j=1}^m v_j L_j$ and using the fact that $(L_j)_{j=1,\dots,m}$ is an orthonormal basis of X_m, we can show using Cauchy-Schwartz inequality that

$$K(X_m) := \sup_{y \in U} \sum_{j=1}^m |L_j(y)|^2. \tag{7.2.18}$$

Although this last definition depends on the basis $\mathcal{B}_L$, one should keep in mind that the definition above shows in the inverse. The quantity $K(X_m)$ only depends on the space X_m and the measure ϱ. The quantity $K(V_m)$ is also a uniform bound on the Froebenius norm of the random matrix $R = (L_j(y)L_k(y))_{1 \leq j,k \leq m}$ and therefore allows to bound the deviation of $\mathbf{G}$ which is its empirical average from its expectation $\mathbf{I}$, based on concentration inequalities for matrix valued random variables.

The main theorem in [32] implies that given $r > 0$ and the number of samples n large enough such that

$$\frac{n}{\log n} \geq \frac{K(X_m)}{\kappa}, \quad \text{with} \quad \kappa := \frac{1 - \ln 2}{2 + 2r} \simeq \frac{0.15}{1 + r}, \tag{7.2.19}$$

then the deviation between $\mathbf{G}$ and $\mathbf{I}$ satisfies the probabilistic estimate

$$\Pr\left\{|||\mathbf{G} - \mathbf{I}||| > \frac{1}{2}\right\} \leq 2n^{-r}. \tag{7.2.20}$$

This estimate implies that with probability at least $1 - 2n^{-r}$ the least square problem is stable: indeed, with at least this probability, one has

$$|||\mathbf{G}^{-1}||| \leq 2 \ \text{ and } \ |||\mathbf{G}||| \leq \frac{3}{2}, \tag{7.2.21}$$

and therefore since $n\mathbf{J}\mathbf{J}^t = \mathbf{G}$, then

$$|||J||| \leq \sqrt{\frac{3}{2}} n^{-1/2}. \tag{7.2.22}$$

It follows from (7.2.10) that

$$\|w\|_{L^2}^2 \leq 6\Big(\frac{1}{n}\sum_{j=1}^{n}|z_j|^2\Big). \tag{7.2.23}$$

This shows that in the noiseless case, we have

$$\|P_m^n u\|_{L^2}^2 \leq 6\|u\|_n^2. \tag{7.2.24}$$

Using this result, the following quasi-optimality results for the truncated least-square approximation are proved in [32] for the noiseless and stochastic noisy models and in [23] for the deterministic noisy model.

- In the noiseless model, if u satisfies a uniform bound L over U, then one has

$$\mathbb{E}(\|u-\tilde{P}_m^n u\|^2) \leq \Big(1+\epsilon(n)\Big)e_m(u)^2 + 8L^2 n^{-r}, \quad \text{where} \quad \epsilon(n) := \frac{4\kappa}{\log(n)} \tag{7.2.25}$$

- In the stochastic noise model, if u satisfies a uniform bound L over U, then

$$\mathbb{E}(\|u-\tilde{w}\|^2) \leq \Big(1+2\epsilon(n)\Big)e_m(u)^2 + 8L^2 n^{-r} + 8\sigma^2\frac{m}{n}, \tag{7.2.26}$$

  where $\sigma^2 := \max_{y\in\Gamma}\mathbb{E}(|z-u(y)|^2|y)$ is the noise level.

- In the deterministic noise model, if u satisfies a uniform bound L

$$\mathbb{E}(\|u-\tilde{w}\|^2) \leq (1+2\epsilon(n))e_m(u)^2 + (8+2\epsilon(n))\|\eta\|^2 + 8L^2 n^{-r}. \tag{7.2.27}$$

It is also desirable to estimate the error between u and its estimator in probability rather than in expectation. In the following we give such an estimate, in the noiseless case and for the non truncated estimator $w = P_m^n u$, however using the best approximation error in the uniform norm

$$e_m(u)_\infty := \inf_{v\in V_m}\|u-v\|_{L^\infty(U)}, \tag{7.2.28}$$

which is obviously larger than $e_m(u)$. The next result was already proven in [66] for the particular case of discrete least-squares on univariate polynomial spaces and for the noiseless model. Here, we treat the more general deterministic noise model

Theorem 7.2.1

Under condition (7.2.19), one has

$$\Pr\Big(\|u-P_m^n u\| \geq (1+\sqrt{2})e_m(u)_\infty\Big) \leq 2n^{-r}. \tag{7.2.29}$$

Proof: Introducing the event set $\Omega_+^n := \{|||\mathbf{G} - \mathbf{I}||| \leq \frac{1}{2}\}$, we know from (7.2.20) that $\Pr(\Omega_+^n) \geq 1 - 2n^{-r}$. Given any draw in Ω_+^n, we have for any $v \in V_m$

$$\|u - P_m^n u\| \leq \|u - v\| + \|v - P_m^n u\| \leq \|u - v\| + \sqrt{2}\|v - P_m^n u\|_n, \tag{7.2.30}$$

where we have used (7.2.16). By the orthogonality identity $\|u - v\|_n^2 = \|u - P_m^n u\|_n^2 + \|P_m^n u - v\|_n^2$, we deduce

$$\|u - P_m^n u\| \leq \|u - v\| + \sqrt{2}\|u - v\|_n \leq (1 + \sqrt{2})\|u - v\|_\infty, \tag{7.2.31}$$

which completes the proof. ■

All these results above lead to the problem of understanding which minimal amount n of sample ensures the validity of condition (7.2.19). In the one-dimensional case $d = 1$, with $X_m = \mathbb{P}_{m-1}$ being the space of polynomials of degree less or equal to $m - 1$ and ϱ being the uniform measure over $[-1, 1]$, we have

$$K(\mathbb{P}_{m-1}) = \Big\| \sum_{j=0}^{m-1} |L_j|^2 \Big\|_{L^\infty([-1,1])} = \sum_{j=0}^{m-1} (2j + 1) = m^2 \tag{7.2.32}$$

where we have used the Legendre polynomials $(L_j)_{0 \leq j \leq m-1}$ normalised in $L^2([-1, 1], d\varrho)$ which form an orthonormal basis of $\mathbb{P}_{m-1}$ and all attain their supremums $\sqrt{2j + 1}$ in 1. Therefore (7.2.19) holds for

$$\frac{n}{\log n} \sim m^2, \tag{7.2.33}$$

meaning that n scales like m^2 up to a logarithmic factor. This relation between n and m was also obtained and used in [66] in order to establish estimates for the discrete least-squares error in probability, however, by different arguments which are more tied to the use of univariate polynomials and the uniform measure. The next section discusses the implications of condition (7.2.19) for the multivariate polynomial spaces $\mathbb{P}_\Lambda$.

7.3 least-squares for multivariate polynomials

In this section, we investigate the implications of the condition (7.2.19) in the setting of multivariate polynomials spaces $\mathbb{P}_\Lambda$. The analysis apply for the domains $[-1, 1]^d$ with $d \geq 1$ and also can be generalized easily for the domain $U := [-1, 1]^{\mathbb{N}}$.We consider here ϱ the uniform measure over $[-1, 1]^d$, i.e.

$$d\varrho := \otimes_{j=1}^d \frac{dy_j}{2}. \tag{7.3.1}$$

When U is considered, the definition of infinite product measure ϱ is explained in §1.2 of Chapter 1. We use the notations $L^2([-1,1]^d, d\varrho)$, $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ of the previous section and denote $\mathcal{F} := \mathbb{N}^d$. Given Λ a finite subset of $\mathcal{F}$, u the unknown real valued function and $(z^i)_{i=1,\cdots,n}$ noiseless or noisy observations of u at the points $(y^i)_{i=1,\cdots,n}$ where the y^i are i.i.d. random variables distributed according to ϱ, we introduce the polynomial discrete least-squares approximation

$$w_\Lambda := \operatorname*{argmin}_{v\in\mathbb{P}_\Lambda} \sum_{i=1}^{n} |z^j - v(y^j)|^2, \tag{7.3.2}$$

In order to study the optimality of the least-squares approximation, we need to investigate the growth of the quantity of interest $K(X_m)$ introduced in (7.2.17) and (7.2.18) with $X_m = \mathbb{P}_\Lambda$. We shall show that under the minimal condition of lower structure of the index set Λ, we have as in the one-dimensional case that $K(\mathbb{P}_\Lambda) \leq (\#\Lambda)^2$.

We introduce $(L_k)_{k\geq 0}$ the univariate Legendre polynomials normalized in $L^2([-1,1], \frac{dt}{2})$ and introduce $(L_\nu)_{\nu\in\mathcal{F}}$ the multivariate Legendre polynomials defined by

$$L_\nu(y) := \prod_{j=1}^{d} L_{\nu_j}(y_j), \quad y \in [-1,1]^d \tag{7.3.3}$$

The family $(L_\nu)_{\nu\in\mathcal{F}}$ is an orthonormal basis of the space $L^2([-1,1]^d, \varrho)$. We have seen in the previous chapters that if Λ is a lower set, then $(L_\nu)_{\nu\in\Lambda}$ is an orthonormal basis of $\mathbb{P}_\Lambda$. Therefore, the multivariate extension of (7.2.18) reads

$$K_L(\mathbb{P}_\Lambda) := \Big\| \sum_{\nu\in\Lambda} |L_\nu| \Big\|_{L^\infty([-1,1]^d)} \tag{7.3.4}$$

Since the univariate Legendre polynomials L_k attain all their maximums $\sqrt{2k+1}$ at 1, then

$$K(\mathbb{P}_\Lambda) = \sum_{\nu\in\Lambda} \|L_\nu\|^2_{L^\infty([-1,1]^d)} = \sum_{\nu\in\Lambda} \prod_{j=1}^{d} (2\nu_j + 1) = K_{0,0}(\Lambda) \tag{7.3.5}$$

where the notation $K_{0,0}(\Lambda)$ is introduced in the appendix (A.4.6) for the same purpose. According to Lemma A.4.1, we have

Lemma 7.3.1

For any finite lower set $\Lambda \subset \mathcal{F}$, the quantity $K(\Lambda)$ satisfies

$$\#(\Lambda) \leq K(\mathbb{P}_\Lambda) \leq (\#(\Lambda))^2. \tag{7.3.6}$$

The previous bound is valid for any lower set independently of its shape. In addition, the inequality is sharp, in the sense that the equality $K(\Lambda) = (\#(\Lambda))^2$ holds for certain

types of lower sets. Indeed, given $\nu \in \mathcal{F}$ supported in $\{1, \cdots, J\}$ and considering a rectangular block $\mathcal{B}_\nu := \{\mu \in \mathcal{F} : \mu \leq \nu\}$, we have

$$K(\mathbb{P}_{\mathcal{B}_\nu}) = \sum_{\mu \leq \nu} \prod_{1 \leq j \leq J} (2\mu_j + 1) = \prod_{1 \leq j \leq J} \sum_{\mu_j \leq \nu_j} (2\mu_j + 1) = \prod_{1 \leq j \leq J} (\nu_j + 1)^2 = (\#\mathcal{B}_\nu)^2. \quad (7.3.7)$$

However, we expect this bound to be pessimistic for lower sets that have shapes very different from rectangles. For instance, we let $k \geq 1$ and consider the simplex

$$\mathcal{S}_k := \{\nu \in \mathbb{N}_0^d : |\nu| \leq k\} \quad (7.3.8)$$

where $|\nu| := \sum_{j=1}^d \nu_j$, associated to the polynomial space $\mathbb{P}_{\mathcal{S}_k}$ of d-variate polynomials of total degree k. By the inequality of arithmetic and geometric means, one has for any $\nu \in S_{k,d}$

$$\prod_{1 \leq j \leq d} (2\nu_j + 1) \leq \Big(\frac{1}{d} \sum_{1 \leq j \leq d} (2\nu_j + 1)\Big)^d = \Big(\frac{2|\nu|}{d} + 1\Big)^d \leq \Big(\frac{2k}{d} + 1\Big)^d. \quad (7.3.9)$$

Therefore

$$K(\mathbb{P}_{\mathcal{S}_k}) \leq \Big(\frac{2k}{d} + 1\Big)^d \#(\mathcal{S}_k). \quad (7.3.10)$$

The quantity $\left(\frac{2k}{d} + 1\right)^d$ is bounded by e^{2k}, hence very small compared to $\#(\mathcal{S}_k) = \binom{d+k}{k}$ for large values of d. On Figure 7.3.1, we provide a comparison between $\#(\mathcal{S}_k)$, $K(\mathbb{P}_{\mathcal{S}_k})$ and $(\#(\mathcal{S}_k))^2$ for various dimensions d.

In light of Lemma 7.3.1, given a finite lower set Λ, if the number of samples n scale like $(\#(\Lambda))^2$ with a logarithmic factor, according to

$$\frac{n}{\log n} \geq \frac{(\#(\Lambda))^2}{\kappa}, \quad \kappa := \frac{1 - \ln 2}{2 + 2r}, \quad (7.3.11)$$

then the stability and approximation results of the previous section hold in the the present setting of multivariate polynomial least-squares approximation.

It is interesting to see if the estimates on the quantity $K(\mathbb{P}_\Lambda)$ can be improved when using other standard probability measures over $[-1, 1]^d$. In what follows, we study this quantity when the measure ϱ is the tensorized Chebychev measure, i.e.

$$d\varrho(y) := \otimes_{j \geq 1} \varrho(y_j) dy_j, \quad \text{with} \quad \varrho(t) := \frac{1}{\pi} \frac{1}{\sqrt{1 - t^2}}. \quad (7.3.12)$$

Using in this case the notation $K_T(\mathbb{P}_\Lambda)$ to denote the quantity of interest $K(\mathbb{P}_\Lambda)$ with the measure ϱ. We have

$$K_T(\mathbb{P}_\Lambda) := \Big\| \sum_{\nu \in \Lambda} |T_\nu|^2 \Big\|_{L^\infty(U)}, \quad (7.3.13)$$

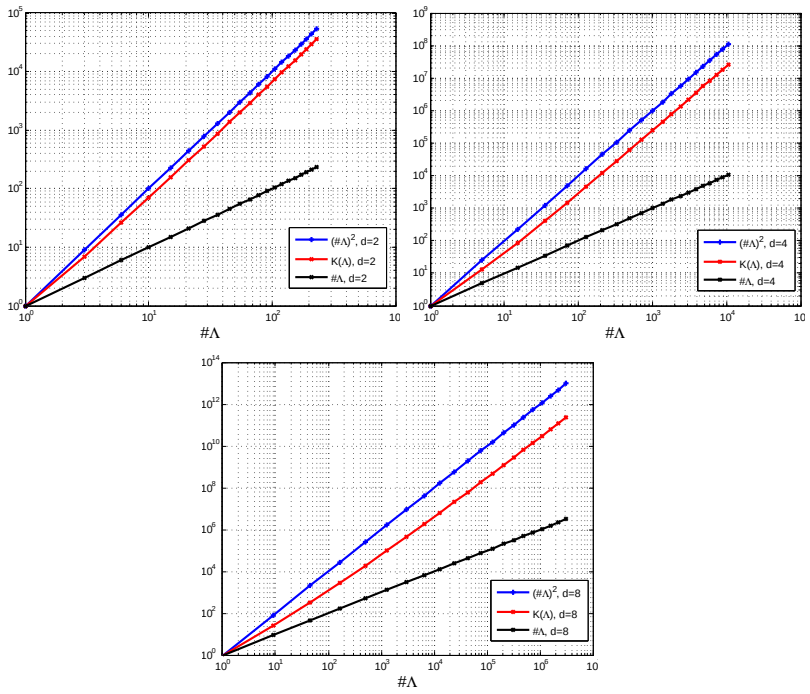

Figure 7.3.1: Comparison between $\#(\mathcal{S}_k)$, $K(\mathbb{P}_{\mathcal{S}_k})$ and $(\#(\mathcal{S}_k))^2$.Left: $d = 2$. Center: $d = 4$. Right: $d = 8$.

where $T_\nu(y) = \prod_{j\geq 1} T_{\nu_j}(y_j)$ are the tensorized Chebyshev polynomials with $(T_k)_{k\geq 0}$ normalized according to

$$\int_{-1}^{1} |T_k(t)|^2 \frac{dt}{\pi\sqrt{1-t^2}} = 1. \tag{7.3.14}$$

It is easily checked that the latter are related to the classical Chebyshev polynomials of the first kind by $T_k(\cos\theta) = \sqrt{2}\cos(k\theta)$ for any $k \geq 1$ and $T_0 = 1$. Therefore, for every $k \geq 2$, the polynomial T_k attains the maximum value $\sqrt{2}$ at 1. It follows that

$$K_T(\mathbb{P}_\Lambda) := \sum_{\nu\in\Lambda} 2^{\#\mathrm{supp}(\nu)} \tag{7.3.15}$$

The quantity of interest $K_T(\mathbb{P}_\Lambda)$ is smaller than the quantity $K(\mathbb{P}_\Lambda)$ defined with the uniform measure and given in (7.3.5). Indeed, for $k \geq 2$, $\|T_k\|_{L^\infty[-1,1]} = \sqrt{2} \leq \sqrt{2k+1}$. Therefore, we obtain

$$K_T(\mathbb{P}_\Lambda) \leq (\#(\Lambda))^2, \tag{7.3.16}$$

A sharper bound can be established by a finer analysis. This is done in the appendix where we have used the notation $K_{-\frac{1}{2},-\frac{1}{2}}(\Lambda)$ to denote the quantity in (7.3.15), see (A.4.7). By Lemma A.4.4, we have

Lemma 7.3.2

For any lower set $\Lambda \subset \mathcal{F}$, the quantity $K_T(\mathbb{P}_\Lambda)$ satisfies

$$K_T(\mathbb{P}\Lambda) \leq (\#(\Lambda))^{\frac{\ln 3}{\ln 2}} \tag{7.3.17}$$

The previous bound is sharp for certain type of lower sets. For instance, if ν is the multi-index such that $\nu_1 = \ldots = \nu_J = 1$ and $\nu_j = 0$ for $j > J$, then

$$K_T(\mathcal{B}_\nu) = \sum_{\mu \leq \nu} 2^{\#\mathrm{supp}(\mu)} = \sum_{\mu \leq \nu} 2^{\mu_1 + \cdots + \mu_J} = \prod_{j=1}^{J} (1+2) = 3^J = (2^J)^{\frac{\ln 3}{\ln 2}\cdot} = (\#\mathcal{B}_\nu)^{\frac{\ln 3}{\ln 2}}. \tag{7.3.18}$$

In the case of finite dimension $d < +\infty$, the following bound can be easily obtained from the result of Lemma 7.3.2:

$$K_T(\Lambda) \leq \min\left\{(\#\Lambda)^{\frac{\ln 3}{\ln 2}}, 2^d \#\Lambda\right\}.$$

Algebraic bounds can also be obtained for the quantity $K(\mathbb{P}_\Lambda)$ when the measure ϱ is any probability measure of the Jacobi type

$$\varrho_{\alpha,\beta}(t) = \frac{(1-t)^\alpha (1+t)^\beta}{W_{\alpha,\beta}}, \quad W_{\alpha,\beta} := \int_{-1}^{1} (1-t)^\alpha (1+t)^\beta dt, \quad \alpha, \beta > -1. \tag{7.3.19}$$

We denote by $(L_k^{\alpha,\beta})_{k\geq 0}$ the corresponding Jacobi polynomials and by $K_{\alpha,\beta}(\mathbb{P}_\Lambda)$ the corresponding quantity. We show in the appendix, formula (A.4.26), that we have a rough bound

$$K_{\alpha,\beta}(\Lambda) \leq (\#(\Lambda))^{\max(2q+1,0)+\gamma}, \quad \gamma = \frac{\ln(C_{\alpha,\beta}^2 + 1)}{\ln 2} \tag{7.3.20}$$

where $C_{\alpha,\beta}$ is the best constant such that

$$\|L_n^{\alpha,\beta}\|_{L^\infty([-1,1])} \leq C_{\alpha,\beta}(n+1)^{\max(q+\frac{1}{2},0)}, \quad q = \max(\alpha,\beta), \quad n \geq 1. \tag{7.3.21}$$

This bound can of course be improved for particular values of α and β as it was done for $\alpha = \beta = 0$ and $\alpha = \beta = -\frac{1}{2}$. We do not pursue this direction. For more general measures, one need to derive algebraic bounds of the infinite norm of the corresponding orthogonal polynomials, then by similar argument as in the appendix derive algebraic bounds on the quantity $K(\mathbb{P}_\Lambda)$ for lower sets Λ.

7.4 Discrete least-squares approximation of Hilbert-valued functions

7.4.1 Stability and accuracy

In sections 7.2 and 7.3, the functions that we propose to approximate using the least-squares method are real-valued. Motivated by the application to parametric PDEs, we investigate the applicability of the least-squares method in the approximation of V-valued functions, with V being any Hilbert space. Similar to §7.2 , we work in the abstract setting of a probability space $([-1,1]^d, \Theta, \varrho)$. We study the least-squares approximation of functions u belonging to the Bochner space

$$L^2(U,V,d\varrho) := \Big\{ u : U \to V, \|u\| := \int_U \|u(y)\|_V^2 d\varrho(y) < +\infty \Big\}. \tag{7.4.1}$$

We have $L^2(U,V,d\varrho) = V \otimes L^2(U,d\varrho)$ and we are interested in the least-squares approximation in spaces of type $V \otimes X_m$ where X_m is an m-dimensional subspace of $L^2(U,d\varrho)$. Given $u \in L^2(U,V,d\varrho)$ an unknown function and $(z^i)_{i=1,\cdots,n}$ noiseless or noisy observations of u (belonging to V) at the points $(y^i)_{i=1,\cdots,n}$ where the y^i are i.i.d. random variables distributed according to ϱ, we consider the discrete least-squares approximation

$$w := \operatorname*{argmin}_{v \in V \otimes X_m} \sum_{i=1}^n \|z^j - v(y^j)\|_V^2. \tag{7.4.2}$$

The purpose of this section is to briefly discuss the extension of the results from §7.2 to the present framework.

Let $\mathcal{B}_L$ be an orthonormal basis of the space X_m with respect to the measure ϱ and consider the matrices $\mathbf{G}$ and $\mathbf{J}$ and the family $\mathcal{B}_\pi \subset X_m$ obtained from the basis $\mathcal{B}_L$ and the points $(y^i)_{i=1,\cdots,n}$ as in §7.2. When the matrix $\mathbf{G}$ is not singular, we claim that the solution to (7.4.2) has the same form as in the real case, namely

$$w = \sum_{j=1}^n z^j \pi_j, \tag{7.4.3}$$

with $z^j \in V$ for all $j = 1, \ldots, n$ are the observation as in the real-valued case. Indeed, from the analysis of §7.2, for any $g \in V$, the real-valued function $w_g := \sum_{j=1}^n \langle z^j, g\rangle \pi_j \in X_m$ is the solution to the least-squares problem

$$w_g = \operatorname*{argmin}_{h \in X_m} \sum_{i=1}^n |\langle z^i, g\rangle - h(y^i)|^2, \tag{7.4.4}$$

which implies the following orthogonality relations over $V \otimes X_m$

$$\sum_{i=1}^{n} \Big\langle \sum_{k=1}^{n} z^k \pi_k(y^i), gL_j(y^i) \Big\rangle = \sum_{i=1}^{n} \langle z^i, gL_j(y^i)\rangle, \quad g \in V,\ j \in \{1, \cdots, m\}, \tag{7.4.5}$$

showing that $\sum_{j=1}^{n} z^j \pi_j$ is the solution to (7.4.2). When the matrix $\mathbf{G}$ is singular, the solution (7.4.2) is not unique and we set by convention $w := 0$.

The explicit formula of the least-squares approximation (7.4.2) being established, we are interested in the stability and accuracy of the approximation. Similarly to the analysis in §7.2, we investigate the comparability over $V \otimes X_m$ of the norm $\|\cdot\|$ and its empirical counterpart $\|\cdot\|_n$ defined by

$$\|v\|_n = \Big(\frac{1}{n}\sum_{j=1}^{n} \|v(y^j)\|_V^2\Big)^{\frac{1}{2}}, \quad v \in L^2(U, V, \varrho). \tag{7.4.6}$$

It is easily checked that given $v := \sum_{j=1}^{m} v_j L_j \in V \otimes X_m$, one has

$$\|v\|_n^2 - \|v\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} (\mathbf{G} - \mathbf{I})_{ij} \langle v_i, v_j\rangle_V = \langle \mathbf{v}, (\mathbf{G} - \mathbf{I})\mathbf{v}\rangle_{V^m}, \tag{7.4.7}$$

where $\mathbf{v} := (v_1, \cdots, v_m)^t \in V^m$ and the matrix-vector product is defined as in the real case. Here the inner product $\langle \cdot, \cdot \rangle_{V^m}$ is the standard inner product over V^m constructed from $\langle \cdot, \cdot \rangle_V$. Note that we have $\|v\| = \|\mathbf{v}\|_{V^m}$. We next observe that if $\mathbf{M}$ is an $m \times m$ real symmetric matrix, then by diagonalizing $\mathbf{M}$ in an orthonormal basis, i.e. $M = \mathbf{P}^t\mathbf{D}\mathbf{P}$ with $\mathbf{P}$ a unitary $m \times m$ matrix and $\mathbf{D}$ diagonal, we easily check that

$$\sup_{\|\mathbf{v}\|_{V^m}=1} |\langle \mathbf{v}, \mathbf{M}\mathbf{v}\rangle_{V^m}| = \sup_{\|\mathbf{v}\|_{V^m}=1} |\langle \mathbf{P}\mathbf{v}, \mathbf{D}\mathbf{P}\mathbf{v}\rangle_{V^m}| = \sup_{\|\mathbf{w}\|_{V^m}=1} |\langle \mathbf{w}, \mathbf{D}\mathbf{w}\rangle_{V^m}| = |||\mathbf{D}||| = |||\mathbf{M}|||, \tag{7.4.8}$$

where $|||\mathbf{M}|||$ is the spectral norm of M. Therefore

$$\|v\|_n^2 - \|v\|^2 \le |||\mathbf{G} - \mathbf{I}|||\, \|v\|^2, \quad v \in V \otimes X_m, \tag{7.4.9}$$

so that similar to the results discussed in §7.3, we find that under the condition (7.2.19), the norm $\|\cdot\|$ and its counterpart $\|\cdot\|_n$ are equivalent over $V \otimes X_m$ with a probability greater than $1 - 2n^{-r}$, with

$$\Big|\|v\|_n^2 - \|v\|^2\Big| \le \frac{1}{2}\|v\|^2, \tag{7.4.10}$$

Similar to real valued functions, we want to compare the accuracy of the least-squares approximation (7.4.2) with the error of best approximation in $L^2([-1,1]^d, V, \varrho)$

$$e_m(u) := \inf_{v \in V \otimes X_m} \|u - v\| = \|u - P_m u\|, \tag{7.4.11}$$

where P_m is the orthogonal projector onto $V \otimes X_m$. We again use the notation $P_m^n u$ for the least-squares solution in the noiseless model. If u satisfies a uniform bound

$$\|u(y)\|_V \leq L, \quad y \in [-1,1]^d, \tag{7.4.12}$$

where L is known, we define the truncated least-squares approximation

$$\tilde{w} = T_L(w), \tag{7.4.13}$$

also denoted by $\tilde{P}_m^n u$ in the noiseless model, where T_L is the truncation operator, now defined as follows

$$T_L(v) := \min\Big(\|v\|, L\Big)\frac{v}{\|v\|} = \begin{cases} v & \text{if } \|v\| \leq L, \\ \\ \frac{L}{\|v\|}v & \text{if } \|v\| > L. \end{cases} \tag{7.4.14}$$

Note that T_L is the projection map onto the closed disc $\{\|v\| \leq L\}$ and is therefore Lipschitz continuous with constant 1. The following counterparts to the results of §7.3 are proved in the exact same way and therefore we only state them:

- Under condition (7.2.19), and if u satisfies a unifom bound $\|u(y)\|_V \leq L$ over $[-1,1]^d$, one has

$$\mathbb{E}(\|u - \tilde{P}_m^n u)\|_V^2) \leq \Big(1 + \epsilon(n)\Big)e_m(u)^2 + 8L^2 n^{-r}. \tag{7.4.15}$$

- In the noisy model, and under the same conditions as above, one has

$$\mathbb{E}(\|u - \tilde{w}\|_V^2) \leq \Big(1 + 2\epsilon(n)\Big)e_m(u)^2 + 8\Big(L^2 n^{-r} + \sigma^2 \frac{m}{n}\Big), \tag{7.4.16}$$

  where $\sigma^2 := \max_{y\in\Gamma} \mathbb{E}(\|z - u(y)\|_X^2|y)$ is the noise level.

- Under condition (7.2.19), one has

$$\Pr\Big(\|u - P_m^n u\|_V \geq (1+\sqrt{2})e_m(u)_\infty\Big) \leq 2n^{-r}, \tag{7.4.17}$$

  where $e_m(u)_\infty = \inf_{v\in V\otimes X_m} \|u - v\|_{L^\infty(U,V)}$.

7.4.2 Application to polynomials approximation of parametric PDEs

As a general example of application,we consider a parametric PDE of the general form (7.1.1) and assume that $\mathcal{D}$ satisfies a (p,ε)-holomorphy assumption for some $p < 1$ with the solution space V a Hilbert space such as $H_0^1(D)$. We assume that y is a random

vector with the uniform probability distribution over U. Given a sequence $(\Lambda_n^L)_{n\geq 1}$ a sequence of nested lower sets yielding the algebraic rate $s^* := \frac{1}{p} - \frac{1}{2}$ in the mean average sense

$$e_{\Lambda_m^L}(u) = \inf_{v \in \mathbb{V}_{\Lambda_m^L}} \|u - v\|_{\mathcal{V}_2} \leq C'(m+1)^{-s^*}, \tag{7.4.18}$$

We propose to compute approximation to u using the least square methods.

From the (p, ε)-holomorphy, the solution satisfies a uniform bound $\|u(y)\|_V \leq L$ over U. We can compute its trunctated least-squares approximation $\tilde{P}_m^n u$ based on n observations $u^i = u(y^i)$ where the y^i are i.i.d. with respect to the uniform measure over U. Combining (7.4.15) and the bound we derived for $K(\Lambda)$ with Legendre polynomials (7.3.6), it follows that

$$\mathbb{E}(\|u - \tilde{P}_m^n u)\|^2) \leq (1 + \epsilon(n))C^2(m+1)^{-2s^*} + 8L^2 n^{-r}, \tag{7.4.19}$$

provided that $\frac{n}{\log n} \geq \frac{(m+1)^2}{\kappa} \geq \frac{m^2}{\kappa}$ with $\kappa := \frac{1-\ln 2}{2+2r}$. In particular, taking $r = s^*$, we obtain that if the number of samples n scales according to

$$\frac{n}{\log n} \geq \frac{(m+1)^2}{\kappa^*} \quad \text{where} \quad \kappa^* = (1 - \ln 2)\frac{p}{p+2}. \tag{7.4.20}$$

then we recover the optimal rate in expectation

$$\sqrt{\mathbb{E}[\|u - \tilde{P}_m^n u)\|_V^2]} \lesssim (m+1)^{-s^*}. \tag{7.4.21}$$

Remark 7.4.1

An analysis of the Chebychev coefficients of u reveals that the same approximation rate as (7.4.21) hold for the L^2 norm with respect to the tensorized Chebychev measure. However, in view of (7.3.17), the condition between m and n is now $\frac{n}{\log n} \geq \frac{m^\beta}{\kappa^*}$ with $\beta := \frac{\log 3}{\log 2}$. It follows that the rate in (7.4.21) can be improved into

$$\sqrt{\mathbb{E}\Big[\|u - \tilde{P}_m^n u)\|^2\Big]} \lesssim (m+1)^{-\frac{\log 3}{\log 2}s}, \tag{7.4.22}$$

if we use samples y^i that are i.i.d with respect to the tensorized Chebychev measure and if we use the L^2 error with respect to this measure.

7.5 Conclusion

In the present chapter the approximation technique based on least-squares with random evaluations has been analyzed. The condition between the number of sampling points and the dimension of the polynomial space, which is necessary to achieve stability

and optimality, has been extended to any lower set of multi-indices identifying the polynomial space, in any dimension of the parameter set, and with the uniform and Chebychev densities. When the density is uniform, this condition requires the number of sampling points to scale as the squares of the dimension of the polynomial space.

Afterwards, this technique has been applied to a class of "inclusion-type" elliptic PDE models with stochastic coefficients, and an exponential convergence rate in expectation has been derived. This estimate clarifies the dependence of the convergence rate on the number of sampling points and on the dimension of the parameter set. Moreover, this estimate establishes a relation between the convergence rate of the least-squares approximation with random evaluations and the convergence rate of the best m-term "exact" L^2 projection.

The numerical tests presented show that the proposed estimate is sharp, when the number of sampling points is chosen according to the condition that ensures stability and optimality. In addition, these results show that, in the aforementioned model class, a linear scaling of the number of sampling points with respect to the dimension seems to be sufficient to ensure the stability of the discrete projection, thus leading to faster convergence rates, although we have no rigourous explaination of this fact.

Appendix A

Jacobi polynomials

A.1 Definitions of Jacobi polynomials.

We consider the family of weight functions $(w_{\alpha,\beta})_{\substack{\alpha>-1\\ \beta>-1}}$ defined over $[-1,1]$ by

$$w_{\alpha,\beta}(t) = (1-t)^{\alpha}(1+t)^{\beta}, \quad t \in [-1,1]. \tag{A.1.1}$$

For the range of values α and β considered, all these weight functions possess finite positive integrals

$$W_{\alpha,\beta} := \int_{-1}^{1} (1-t)^{\alpha}(1+t)^{\beta} dt = 2^{\alpha+\beta+1} \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}. \tag{A.1.2}$$

To see this, one make the change of variable $s = \frac{1+t}{2}$ and get

$$W_{\alpha,\beta} = 2^{\alpha+\beta+1} \int_{0}^{1} (1-s)^{\alpha} s^{\beta} ds = 2^{\alpha+\beta+1} B(\beta+1, \alpha+1), \tag{A.1.3}$$

where B is the beta function and remark that $\beta+1$ and $\alpha+1$ are both positive, hence the value of $B(\beta+1,\alpha+1)$ is known from the classical properties of beta function. We denote by $(p_n^{\alpha,\beta})_{n\geq 0}$ the family of the so-called Jacobi polynomials associated with α and β. For any given $\alpha > -1$ and $\beta > -1$, the family $(p_n^{\alpha,\beta})_{n\geq 0}$ is orthonormal with respect to $w_{\alpha,\beta}$. We denote by $(P_n^{\alpha,\beta})_{n\geq 0}$ the family of the so-called orthogonal Jacobi polynomials associated with α and β. Theses polynomials are orthogonal with respect to $w_{\alpha,\beta}$ but normalised according to

$$P_n^{\alpha,\beta}(1) = \binom{n+\alpha}{n} := \frac{\Gamma(n+\alpha+1)}{\Gamma(n+1)\Gamma(\alpha+1)}, \tag{A.1.4}$$

There exists an immense literature dealing with the study of theses polynomials, e.g. [35, 20, 79]. The notation we have used for the family of polynomials above is borrowed to [35]. We recall without proof the classical results on Jacobi polynomials and we establish some others results that we have used for various purposes within the thesis manuscript.

The polynomials $P_n^{\alpha,\beta}$ have closed formulas, known as Rodrigues formulas, namely

$$P_n^{\alpha,\beta}(t) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\Big((1-x)^{n+\alpha}(1+x)^{n+\beta}\Big). \tag{A.1.5}$$

This result is given and proved in [35, Theorem 10.3.1].

We consider the family of probabilistic weight functions $(\varrho_{\alpha,\beta})_{\substack{\alpha>-1\\ \beta>-1}}$ defined over $[-1,1]$ by

$$\varrho_{\alpha,\beta} := \frac{w_{\alpha,\beta}}{W_{\alpha,\beta}}, \tag{A.1.6}$$

and introduce the families $(L_n^{\alpha,\beta})_{n\geq 0}$ of "*probabilistic Jacobi polynomials*" in $[-1,1]$, i.e. the polynomials $(L_n^{\alpha,\beta})_{n\geq 0}$ are orthonormal with respect to $\varrho_{\alpha,\beta}$. For any given $\alpha > -1$ and $\beta > -1$, the family $(L_n^{\alpha,\beta})_{n\geq 0}$ form an orthonormal basis of the Hilbert space $L^2([-1,1], d\varrho_{\alpha,\beta})$. It is immediate that

$$L_n^{\alpha,\beta} = \sqrt{W_{\alpha,\beta}}p_n^{\alpha,\beta}, \quad n \geq 0. \tag{A.1.7}$$

We recall also the relation between the polynomials $p_n^{\alpha,\beta}$ and $P_n^{\alpha,\beta}$ established in in [35, Corollary 10.3.6]

$$p_n^{\alpha,\beta} = M_n^{\alpha,\beta}P_n^{\alpha,\beta}, \quad n \geq 0. \tag{A.1.8}$$

where for any $n \geq 0$

$$M_n^{\alpha,\beta} := \left\{\frac{(2n+\alpha+\beta+1)}{2^{\alpha+\beta+1}}\frac{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}\right\}^{1/2}. \tag{A.1.9}$$

The three last equalities combined imply that

$$L_n^{\alpha,\beta} = N_n^{\alpha,\beta}P_n^{\alpha,\beta}, \quad n \geq 0 \tag{A.1.10}$$

where for any $n \geq 0$, $N_n^{\alpha,\beta} = M_n^{\alpha,\beta}\sqrt{W_{\alpha,\beta}}$. Using the formula (A.1.2) and rearranging the various terms to our convenience, we infer the explicit formulas

$$N_n^{\alpha,\beta} = \left\{\frac{(2n+\alpha+\beta+1)}{(\alpha+\beta+1)}\frac{\binom{n+\alpha+\beta}{n}}{\binom{n+\alpha}{n}\binom{n+\beta}{n}}\right\}^{1/2}. \tag{A.1.11}$$

with 1 instead of $(\alpha+\beta+1)$ in the case this sum is equal to 0. Let us note that from theirs definitions, the polynomials $L_0^{\alpha,\beta}$ are all constant equal to 1 for any $\alpha, \beta > -1$. This can also be checked using that $N_0^{\alpha,\beta} = 1$ and $P_0^{\alpha,\beta}$ is constant equal to $P_0^{\alpha,\beta}(1) = 1$ given by (A.1.4).

A.2 Supremum norms of orthonormal Jacobi polynomials

We are interested in bounding the supremum norm over $[-1,1]$ of the polynomials $L_n^{\alpha,\beta}$. To this end, we first recall the exact values of these norms for the polynomials $P_n^{\alpha,\beta}$ given in [79, Theorem 7.32.1] .

Theorem A.2.1

Given $\alpha > -1$, $\beta > -1$, $q = \max(\alpha,\beta)$ and $t_0 = \frac{\beta-\alpha}{\alpha+\beta+1}$, one has

$$\|P_n^{\alpha,\beta}\|_{L^\infty([-1,1])} = \begin{cases} \binom{n+q}{n} \sim n^q & \text{if} \quad q \geq -\frac{1}{2}, \\ |P_n^{\alpha,\beta}(t')| \sim n^{-\frac{1}{2}} & \text{if} \quad q < -\frac{1}{2}, \end{cases} \tag{A.2.1}$$

where t' is nearest point to t_0 where a maximum is attained and $\sim$ the order of magnitude as $n \to \infty$.

The previous theorem combined with (A.2.8) provides the exact values of $\|L_n^{\alpha,\beta}\|_{L^\infty}$ for the classical Jacobi polynomials. More precisely, for Legendre polynomials ($\alpha = \beta = 0$), Tchybeshev polynomials of the first and second kind ($\alpha = \beta = \pm\frac{1}{2}$), a straightforward application of the previous lemma taking into account the normalisation coefficient in (A.2.8) shows that for any $n \geq 1$

$$\|L_n^{0,0}\|_{L^\infty} = \sqrt{2n+1}, \qquad \|L_n^{-\frac{1}{2},-\frac{1}{2}}\|_{L^\infty} = \sqrt{2}, \qquad \|L_n^{\frac{1}{2},\frac{1}{2}}\|_{L^\infty} = n+1. \tag{A.2.2}$$

Also, by direct computations, one has

$$\|L_n^{\frac{1}{2},-\frac{1}{2}}\|_{L^\infty} = \|L_n^{-\frac{1}{2},\frac{1}{2}}\|_{L^\infty} = 2n+1. \tag{A.2.3}$$

We now turn our attention to arbitrary values of $\alpha, \beta > -1$. First, by Sterling equivalents of the gamma function, we have $\binom{n+t}{n} \sim n^t$ for any $t \in \mathbb{R}$. Therefore normalisation coefficient in (A.2.8) is of order $\sqrt{n}$ so that in view of the previous lemma

$$\|L_n^{\alpha,\beta}\|_{L^\infty} \lesssim n^{\max(q,-\frac{1}{2})+\frac{1}{2}}, \quad q = \max(\alpha,\beta), \quad n \geq 1. \tag{A.2.4}$$

We recall that $\sim$ stands here for the order of magnitude as $n \to \infty$ and not the equivalence an $n \to \infty$.

Our analysis of the quantities $\|L_n^{\alpha,\beta}\|_{L^\infty}$ is motivated by the convergence of Jacobi series in chapters I and II and by the study of the growth of sums of square supremum of Jacobi polynomials from chapter VII. The analysis therein shows that one need precise bound of $\|L_n^{\alpha,\beta}\|_{L^\infty}$ for any $n \geq 1$, note only their order of magnitude as $n \to +\infty$. We are therefore interested in finding sharp constants $C > 0, c > -1$ such that

$$\|L_n^{\alpha,\beta}\|_{L^\infty} \leq C(n+c)^{\max(q,-\frac{1}{2})+\frac{1}{2}}, \quad q = \max(\alpha,\beta), \quad n \geq 1. \tag{A.2.5}$$

In the case $q = \max(\alpha, \beta) \geq -\frac{1}{2}$, the values $\|L_n^{\alpha,\beta}\|_{L^\infty}$ are explicit and sharp upper bounds can be obtained using Sterling type inequalities. In the case when $q < -\frac{1}{2}$, the second bound in Lemma A.2.1 was obtained using the result [79, formula 8.21.10] which shows that $\|P_n^{\alpha,\beta}\|_{L^\infty} \leq \frac{1}{\sqrt{\pi n}} + \mathcal{O}(n^{-\frac{3}{2}})$, however the constant in $\mathcal{O}$ is not studied. Rather than striving for at most generality, we only study the case where $\alpha = \beta \in]-1, \infty[$ which corresponds to the so-called ultra-spherical polynomials. The key points in the analysis of $\|L_n^{\alpha,\beta}\|_{L^\infty}$ are the Sterling inequalities. We shall rely on the following from the paper [10],

$$t^t e^{-t}\sqrt{2\pi(t+a)} < \Gamma(t+1) < t^t e^{-t}\sqrt{2\pi(t+b)}, \qquad t \geq 0, \tag{A.2.6}$$

with the best possible constants $a = \frac{1}{6} = 0.166\ldots$ and $b = \frac{e^2}{2\pi} - 1 = 0.176\ldots$. The previous inequality combined with function study and the classical inequality $(1 + t/x)^x \leq e^t$ for any $t, x \geq 0$ implies the following growth result

$$\frac{\Gamma(x+t+1)}{\Gamma(x+1)} \leq (x+t)^t \sqrt{\frac{1+t+b}{1+a}}, \qquad t \geq 0, \quad x \geq 1. \tag{A.2.7}$$

Combining (A.2.8) and Theorem A.2.1, we have for any $\lambda > -\frac{1}{2}$ and any $n \geq 0$,

$$\|L_n^{\lambda,\lambda}\|_{L^\infty} = \left\{ \frac{2n+2\lambda+1}{2\lambda+1} \binom{n+2\lambda}{n} \right\}^{1/2} = \left\{ \frac{2n+2\lambda+1}{\Gamma(2\lambda+2)} \frac{\Gamma(n+2\lambda+1)}{\Gamma(n+1)} \right\}^{1/2}. \tag{A.2.8}$$

We have then the following bounds

Theorem A.2.2

Let $\lambda \in]-\frac{1}{2}, 0]$, for any $n \geq 1$

$$\|L_n^{\lambda,\lambda}\|_{L^\infty} \leq \sqrt{2/\pi} C_1 \sqrt{2n+1}(n+1)^\lambda, \qquad C_1 := 1.2 \tag{A.2.9}$$

Let $\lambda \in [0, \frac{1}{2}]$, for any $n \geq 1$

$$\|L_n^{\lambda,\lambda}\|_{L^\infty} \leq \sqrt{2} C_2 (n+1)^{\lambda+\frac{1}{2}}, \qquad C_2 := 1.002 \tag{A.2.10}$$

Proof: If $\lambda \in]-\frac{1}{2}, 0]$, then $2\lambda + 1 > 0$, therefore by (A.2.7) applied with $x = n$ and $t = 2\lambda + 1$, we deduce

$$\frac{\Gamma(n+2\lambda+1)}{\Gamma(n+1)} \leq (n+2\lambda+1)^{2\lambda} \sqrt{\frac{1+2\lambda+1+b}{1+a}} \leq (C_1)^2 (n+1)^{2\lambda}$$

where the second inequality is justified by $\lambda \leq 0$ and a function study in the variable $\lambda \in]-\frac{1}{2}, 0]$ on the part not depending on n. Since $\lambda \in]-\frac{1}{2}, 0]$, then $2n+2\lambda+1 \leq 2n+1$ and $\Gamma(2\lambda+2) \geq 1$, because the Gamma function in monotone decreasing over]0,1]. the proof of the first part is complete.

Now if $\lambda \geq 0$, then by the growth inequality (A.2.7) applied this time with $x = n$ and $t = 2\lambda \geq 0$, we obtain

$$\frac{\Gamma(n+2\lambda+1)}{\Gamma(n+1)} \leq (n+2\lambda)^{2\lambda} \sqrt{\frac{1+2\lambda+b}{1+a}} \leq (C_2)^2 (n+1)^{2\lambda}$$

where the second inequality follows by function study of the part not depending on n with $\lambda \in [0, \frac{1}{2}]$. Since $2\lambda + 2 \in [2, 3]$ then $\Gamma(2\lambda+2) \geq 1$ and $2n + 2\lambda + 1 \leq 2(n+1)$, which implies the second part in the lemma. ■

In view of (A.2.4), in the case $q < -\frac{1}{2}$, the norms $\|L_n^{\alpha,\beta}\|_{L^\infty[-1,1]}$ stay bounded as n grows. We provide a bound in the case α, β.

Theorem A.2.3

Let $\lambda \in]-1, -\frac{1}{2}[$, we have

$$\|L_n^{\lambda,\lambda}\|_{L^\infty} \leq \sqrt{2C_0} \frac{\sqrt{\Gamma(2\lambda+2)}}{2^{\lambda+\frac{1}{2}} \Gamma(\frac{2\lambda+1}{2}+1)} \tag{A.2.11}$$

with $C_0 = \sqrt{\frac{1+b}{1+a}} \frac{1+2b}{1+2a} \sim 1.018$. The bound is sharp for $n \to \infty$.

Proof: First, we have

$$P_n^{\lambda,\lambda} = \frac{\binom{n+\lambda}{n}}{\binom{n+2\lambda}{n}} P_n^{(\lambda+\frac{1}{2})} \tag{A.2.12}$$

where $P_n^{(\lambda)}$ denote the ultra spherical polynomials as defined in [79] in formula (4.7.1). Now using (A.2.8),taking into account $\lambda \in]-1, -\frac{1}{2}[$, we obtain

$$N_n^{\lambda,\lambda} = \left\{ \frac{(2n+2\lambda+1)}{(2\lambda+1)} \right\}^{1/2} \left\{ \binom{n+2\lambda}{n} \right\}^{1/2} \frac{1}{\binom{n+\lambda}{n}} \tag{A.2.13}$$

Consequently by (A.1.10), we infer

$$\|L_n^{\lambda,\lambda}\| = \left\{ \frac{(2n+2\lambda+1)}{(2\lambda+1)} \right\}^{1/2} \left\{ \binom{n+2\lambda}{n} \right\}^{-1/2} \|P_n^{(\lambda+\frac{1}{2})}\|_{L^\infty}.$$

Bounds of $\|P_n^{(\lambda+\frac{1}{2})}\|_{L^\infty}$ are given in [79, Theorem 7.33.1]. The inspection of the proof of the theorem shows the following, we introduce k and p defined by $k = \frac{n}{2}$, $p = n+2\lambda+1$ if n even and $k = \frac{n-1}{2}$, $p = n$ if n odd, then we have for any $n \geq 1$

$$\|P_n^{(\lambda+\frac{1}{2})}\|_{L^\infty} \leq \left| \binom{k + \frac{2\lambda+1}{2}}{k} \right| \frac{1}{\sqrt{p}} \frac{|2\lambda+1|}{\sqrt{n+2\lambda+1}},$$

$$\|P_n^{(\lambda+\frac{1}{2})}\|_{L^\infty} \leq \begin{cases} \frac{|2\lambda+1|}{(n+2\lambda+1)}\left|\binom{k+\frac{2\lambda+1}{2}}{k}\right| & \text{if } n \text{ even}, \\ \\ \frac{|2\lambda+1|}{\sqrt{n(n+2\lambda+1)}}\left|\binom{k+\frac{2\lambda+1}{2}}{k}\right| & \text{if } n \text{ odd}, \end{cases}$$

with an inequality in the case where n is even and the bound is sharp in the case $n \to \infty$ odd. Up to rearranging the different terms accordingly, we have the unified formula

$$\|L_n^{\lambda,\lambda}\| \leq \frac{\sqrt{\Gamma(2\lambda+2)}}{\Gamma(\frac{2\lambda+1}{2}+1)}\left\{\frac{2n+2\lambda+1}{n+2\lambda+1}\right\}^{1/2}\frac{1}{\sqrt{p}}\left\{\frac{C_k}{C_{k+2\lambda+1}}\right\}^{\frac{1}{2}},$$

where

$$C_t := \frac{\Gamma(t+1)}{\Gamma(\frac{t}{2}+1)\Gamma(\frac{t}{2}+1)}, \quad t \geq 0. \tag{A.2.14}$$

Using sterling inequality (A.2.6) above, we have for any $t \geq 0$ that

$$\frac{\sqrt{t+a}}{(\frac{t}{2}+b)} \leq \sqrt{2\pi}\frac{C_t}{2^t} \leq \frac{\sqrt{t+b}}{(\frac{t}{2}+a)},$$

therefore, for any $t \geq 1$, we have

$$2^{2\lambda+1}\frac{C_t}{C_{t+2\lambda+1}} \leq \frac{\sqrt{t+b}}{(t+2a)}\frac{((t+2\lambda+1)+2b)}{\sqrt{(t+2\lambda+1)+a}}$$

The term to the right considered as a function of λ is increasing in $]-1,-\frac{1}{2}[$, then considered as a function of $t \geq 1$ is decreasing, we deduce that it is always smaller than $C = \sqrt{\frac{1+b}{1+a}}\frac{1+2b}{1+2a}$ and it is even smaller that 1 for values of λ far from $-\frac{1}{2}$. This shows that the bound given in the Lemma is valid for any $n \geq 1$. The bound is sharp since the bound for $\|P_n^{(\lambda+\frac{1}{2})}\|_{L^\infty}$ is sharp and the equivalences used are also sharp. ■

We now turn to the case $q = \max(\alpha,\beta) \geq -\frac{1}{2}$. We provide bounds for $\|L_n^{\alpha,\beta}\|_{L^\infty}$ for any values α and β with the sharpest bounds possible for the ultra spherical case for which computation can be straightforward as we have just noted in the previous lemma.

A.3 Jacobi polynomials of the second kind on Bernstein ellipses

Having defined the Jacobi polynomials of the first kind, we now turn our attention to the function $Q_n^{\alpha,\beta}$ defined by

$$Q_n^{\alpha,\beta}(\xi) := \frac{1}{2(1-\xi)^\alpha(1+\xi)^\beta}\int_{-1}^{1}\frac{P_n^{\alpha,\beta}(t)}{\xi-t}w_{\alpha,\beta}(t)dt, \quad \xi \notin [-1,1], \quad n \geq 0. \tag{A.3.1}$$

These functions are the so-called Jacobi function of the second kind and are also denoted $Q_n^{\alpha,\beta}$ and extensively studied in the literature, see [35, 79]. Here, we are mostly interested in bounding these functions for ξ belonging to the Berstein ellipses $\mathcal{E}_s$. These ellipses are the closed curve defined on the complex plane by

$$\mathcal{E}_s := \Big\{ \frac{z+z^{-1}}{2} : |z| = s \Big\}, \quad s > 1. \tag{A.3.2}$$

The analysis for bounding the Jacobi function of second kind has already been done for the particular Legendre case ($\alpha = \beta = 0$). It has indeed been shown that

$$\Big| Q_n^{0,0}\Big(\frac{z+z^{-1}}{2} \Big) \Big| \leq \frac{\pi}{|z|-1} |z|^{-n}, \quad |z| > 1. \tag{A.3.3}$$

This result is given at the bottom of page 313 in [35] and is a direct application of [35, Lemma 12.4.6]. We are interested in finding similar bounds for arbitrary value of $\alpha > -1$ and $\beta > -1$.

For the purpose of our analysis in Chapter I and II, we are only interested in bounding the integral part in (A.3.1) for the polynomials $L_n^{\alpha,\beta}$ and with the probability measure $\varrho_{\alpha,\beta}$. From this point on, we drop the non integral part and normalise accordingly, yet we keep the same notation. That is

$$Q_n^{\alpha,\beta}(\xi) := \int_{-1}^{1} \frac{L_n^{\alpha,\beta}(t)}{\xi - t} \varrho_{\alpha,\beta}(t) dt, \quad \xi \notin [-1,1], \quad n \geq 0. \tag{A.3.4}$$

The reader can recover the classical Jacobi function of the second kind by multiplying the previous functions by

$$\frac{1}{2(1-\xi)^\alpha (1+\xi)^\beta} \frac{W_{\alpha,\beta}}{N_n^{\alpha,\beta}} = \frac{1}{2(1-\xi)^\alpha (1+\xi)^\beta} \frac{\sqrt{W_{\alpha,\beta}}}{M_n^{\alpha,\beta}} \tag{A.3.5}$$

where $W_{\alpha,\beta}$, $M_n^{\alpha,\beta}$ and $N_n^{\alpha,\beta}$ are introduced in the previous section. To keep our document self contained, we give and prove the result given in [35, Lemma 12.4.6] for Legendre polynomials, for arbitrary values of $\alpha, \beta > -1$.

Lemma A.3.1

Let $\alpha, \beta > -1$, $z \in \mathbb{C}$ such that $|z| > 1$ and $\xi = \frac{z+z^{-1}}{2}$. We have

$$Q_n^{\alpha,\beta}(\xi) = \sum_{m=n+1}^{\infty} \frac{\sigma_{n,m}}{z^m} \tag{A.3.6}$$

with

$$\sigma_{n,m} := 2\int_{-1}^{1} L_n^{\alpha,\beta}(t)U_{m-1}(t)\varrho_{\alpha,\beta}(t)dt = 2\int_{0}^{\pi} L_n^{\alpha,\beta}(\cos\theta)\sin(m\theta)\varrho_{\alpha,\beta}(\cos\theta)d\theta \tag{A.3.7}$$

Proof: Let $z \in \mathbb{C}$ with $|z| > 1$ and $\xi = \frac{z+z^{-1}}{2}$. It is easily checked that $\xi \notin [-1,1]$ and $\xi - t = \frac{z}{2}\Big(1 - \frac{2t}{z} + \frac{1}{z^2}\Big)$ for any $t \in [-1,1]$, therefore

$$Q_n^{\alpha,\beta}(\xi) = \frac{2}{z}\int_{-1}^{1} \frac{L_n^{\alpha,\beta}(t)}{1-\frac{2t}{z}+\frac{1}{z^2}}\varrho_{\alpha,\beta}(t)dt = 2\int_{0}^{\pi} \frac{\frac{\sin\theta}{z}}{1-\frac{2\cos(\theta)}{z}+\frac{1}{z^2}} L_n^{\alpha,\beta}(\cos\theta)\varrho_{\alpha,\beta}(\cos\theta)d\theta, \tag{A.3.8}$$

where we have used the change of variable $t = \cos(\theta)$. Since $|z| > 1$, we have the expansion

$$\begin{aligned}
\frac{\sin\theta}{z}\frac{1}{\Big(1-\frac{2\cos(\theta)}{z}+\frac{1}{z^2}\Big)} &= \frac{1}{2i}\Big(\frac{e^{i\theta}}{z} - \frac{e^{-i\theta}}{z}\Big)\frac{1}{\Big(1-\frac{e^{i\theta}}{z}\Big)\Big(1-\frac{e^{-i\theta}}{z}\Big)} \\
&= \frac{1}{2i}\left(\frac{\frac{e^{i\theta}}{z}}{1-\frac{e^{i\theta}}{z}} - \frac{\frac{e^{-i\theta}}{z}}{1-\frac{e^{-i\theta}}{z}}\right) \\
&= \frac{1}{2i}\sum_{m=1}^{\infty}\Big[\frac{e^{mi\theta}}{z^m} - \frac{e^{-im\theta}}{z^m}\Big] \\
&= \sum_{m=1}^{\infty}\frac{\sin(m\theta)}{z^m}
\end{aligned}$$

Multiplying the last equality by $2L_n^{\alpha,\beta}(\cos\theta)\varrho_{\alpha,\beta}(\cos\theta)$ and integrating between 0 and π, we infer

$$Q_n^{\alpha,\beta}(\xi) = 2\int_{0}^{\pi}\sum_{m=1}^{\infty}\frac{\sin(m\theta)}{z^m}L_n^{\alpha,\beta}(\cos\theta)\varrho_{\alpha,\beta}(\cos\theta)d\theta = 2\int_{-1}^{1}\sum_{m=1}^{\infty}\frac{U_{m-1}(t)}{z^m}L_n^{\alpha,\beta}(t)\varrho_{\alpha,\beta}(t)dt$$

We only need to interchange the integral and the sum and show that $\sigma_{n,m} = 0$ for $m \le n$. Using Cauchy-Schwartz inequality, the bound $\|U_{m-1}\|_{L^\infty[-1,1]} \le m$ and the fact that $\varrho_{\alpha,\beta}$ is a probability measure, we infer that

$$\Big|\int_{-1}^{1} L_n^{\alpha,\beta}(t)U_{m-1}(t)\varrho_{\alpha,\beta}(t)dt\Big| \le \Big(\int_{-1}^{1} L_n^{\alpha,\beta}(t)^2\varrho_{\alpha,\beta}(t)dt\Big)^{\frac{1}{2}}\Big(\int_{-1}^{1} U_{m-1}(t)^2\varrho_{\alpha,\beta}(t)dt\Big)^{\frac{1}{2}} \le m. \tag{A.3.9}$$

The series $\sum \frac{m}{z^m}$ converges since $|z| > 1$ and we may interchange the sum and the integral. This yields

$$Q_n^{\alpha,\beta}(\xi) = \sum_{m=1}^{\infty}\frac{\sigma_{n,m}}{z^m}, \quad \text{with} \quad \sigma_{n,m} = 2\int_{-1}^{1} L_n^{\alpha,\beta}(t)U_{m-1}(t)\varrho_{\alpha,\beta}(t)ds.$$

Since $L_n^{\alpha,\beta}$ is orthogonal to $\mathbb{P}_{n-1}$ with respect to the measure $\varrho_{\alpha,\beta}(t)dt$, then $\sigma_{m,n} = 0$ for any m such that $m - 1 \leq n - 1$, which finishes the proof. ■

The previous lemma has implication on the growth of $Q_n^{\alpha,\beta}$ on the ellipses $\mathcal{E}_s$. We have

Corollary A.3.2

Let $\alpha, \beta > -1$ and $s > 1$. For any $n \geq 1$

$$\sup_{\xi \in \mathcal{E}_s} |Q_n^{\alpha,\beta}(\xi)| = \sup_{z \in \mathcal{U}_s} \left| Q_n^{\alpha,\beta}\Big(\frac{z + z^{-1}}{2}\Big) \right| \leq 2(n+1)\frac{s}{(s-1)^2} s^{-n}. \tag{A.3.10}$$

Proof: We have shown in (A.3.9) that $|\sigma_{m,n}| \leq 2m$ for any $m, n \geq 1$. Now, using classical arguments, we have

$$\sum_{n+1}^{\infty} 2mt^m = 2t\Big(\frac{(n+1)t^n}{1-t} + \frac{t^{n+1}}{(1-t)^2}\Big) \leq 2(n+1)\frac{t^{n+1}}{(t-1)^2}$$

for any $t \in [0, 1[$. Therefore, the triangular inequality applied to (A.3.6) and the previous inequality with $t = \frac{1}{|z|} = \frac{1}{s}$ for z varying in $\mathcal{U}_s$ implies the result. ■

The upper bound in (A.3.10) is quite pessimistic. It is based on the bound $|\sigma_{m,n}| \leq 2m$, which has been obtained in (A.3.9) using Cauchy-Schwartz inequality and the bound $\|U_{m-1}\|_{L^\infty[-1,1]} \leq m$. For example for Legendre polynomials $\alpha = \beta = 0$, using that $\|L_n^{\alpha,\beta}\|_{L^\infty[-1,1]} = \sqrt{2n+1}$ and $\varrho_{\alpha,\beta} \equiv \frac{1}{2}$, the second part in (A.3.7) implies $|\sigma_{n,m}| \leq \pi\sqrt{2n+1}$, which combined with triangular inequality applied to (A.3.6) implies the bound $\pi\sqrt{2n+1}\frac{s^{-n}}{s-1}$ in (A.3.10). The same argument allow us to improve the bound (A.3.10) for $\alpha, \beta \in [0, \frac{1}{2}]$, because for such values $\|L_n^{\alpha,\beta}\|_{L^\infty[-1,1]} \lesssim n^{\frac{1}{2}+\max(\alpha,\beta)}$ and $\varrho_{\alpha,\beta} \lesssim 1$.

We propose to show that the bound in (A.3.10) can be improved for a large class of values $\alpha, \beta > -1$. First, we provide the result for $\alpha = \beta \in \{0, \pm\frac{1}{2}\}$, corresponding to the classical Legendre and Tchybeshev polynomials of first and second kind, for which closed formulas for $Q_n^{\alpha,\beta}$ can be obtained. Then we treat the case $\alpha, \beta > 0$ using elementary arguments. Finally, we treat the case $\alpha, \beta \in]-\frac{1}{2}, +\infty[$ using finer arguments.

Theorem A.3.3

Let $s > 1$. For any $n \geq 0$, we have

$$\sup_{\xi \in \mathcal{E}_s} |Q_n^{0,0}(\xi)| \leq \frac{\sqrt{2\pi}}{s-1} s^{-n}.$$

Proof: The Jacobi functions Q_n associated with Legendre polynomials are given in [79] using series of type (A.3.6). Taking into account the normalisation factor in (A.3.5) which is equal to $\frac{1}{\sqrt{2n+1}}$, we have according to [79, formula 4.9.13], for $z \in \mathbb{C}$ with $|z| > 1$, $\xi = \frac{z+z^{-1}}{2}$ and for any $n \geq 1$

$$Q_n^{0,0}(\xi) = \frac{2}{\sqrt{2n+1}} \frac{4^n (n!)^2}{(2n)!} \sum_{j=0}^{\infty} \frac{f_j}{z^{n+2j+1}},$$

where the f_j are defined by

$$f_0 := 1, \quad f_j := g_j \binom{n+j}{n} \Big/ \binom{n+j+\frac{1}{2}}{n+\frac{1}{2}} \quad \text{with} \quad g_j := \frac{1}{4^m} \binom{2m}{m}.$$

By Sterling type equivalence formulas, we have $\frac{1}{\sqrt{2n+1}} \frac{4^n (n!)^2}{(2n)!} \leq \sqrt{\frac{\pi}{2}}$ for any $n \geq 1$. Using the bound $|g_j| \leq 1$ for any $j \geq 0$, we infer $|f_j| \leq 1$ for any $j \geq 0$. Therefore

$$|Q_n^{0,0}(\xi)| \leq \frac{\sqrt{2\pi}}{|z|^{n+1}} \sum_{j=0}^{\infty} \frac{1}{|z|^{2j}} = \sqrt{2\pi} \frac{|z|}{(|z|-1)(|z|+1)} |z|^{-n},$$

which implies the result. Sharper bounds can be obtained for g_j and f_j using Sterling inequalities yielding $f_j \lesssim \frac{\sqrt{n}}{j}$ for any $j \geq 1$ which yields to the replacement of $\frac{1}{|z|-1}$ by $\mathcal{O}(\log(|z|-1))$ in the formula. ∎

Lemma A.3.4

Let $z \in \mathbb{C}$ with $|z| > 1$ and $\xi = \frac{z+z^{-1}}{2}$. For any $n \geq 1$, $Q_n^{\frac{1}{2},\frac{1}{2}}(\xi) = 2z^{-n-1}$ and $Q_n^{-\frac{1}{2},-\frac{1}{2}}(\xi) = \frac{2\sqrt{2}z}{z^2-1} z^{-n}$. In particular, for any $s > 1$

$$\max_{\xi \in \mathcal{E}_s} |Q_n^{\frac{1}{2},\frac{1}{2}}(\xi)| = 2s^{-n-1}, \quad \text{and} \quad Q_n^{-\frac{1}{2},-\frac{1}{2}}(\xi) \leq \frac{2\sqrt{2}s}{s^2-1} s^{-n} \tag{A.3.11}$$

Proof: Let $(T_n)_{n\geq 0}$ and $(U_n)_{n\geq 0}$ be the Tchybeshev polynomials of the first and second kind defined by

$$U_0 = T_0 \equiv 1 \quad \text{and} \quad \cos(n\theta) = T_n(\cos\theta), \quad \sin((n+1)\theta) = U_n(\cos\theta)\sin\theta, \quad n \geq 1.$$

These two families are orthogonal with respect to the measures $\frac{dt}{\sqrt{1-t^2}}$ and $\sqrt{1-t^2}dt$ respectively with

$$\int_{-1}^{1} T_n^2(t) \frac{1}{\sqrt{1-t^2}} dt = \int_{-1}^{1} U_n^2(t) \sqrt{1-t^2} dt = \frac{\pi}{2}, \qquad n \geq 1,$$

Using (A.1.2) and the value $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ or elementary trigonometric computation, we have $\varrho_{\frac{1}{2},\frac{1}{2}} \equiv \frac{2}{\pi}\sqrt{1-t^2}$ and $\varrho_{-\frac{1}{2},-\frac{1}{2}} \equiv \frac{1}{\pi\sqrt{1-t^2}}$, so that for any $n \geq 1$, $L_n^{\frac{1}{2},\frac{1}{2}} = U_n$ and $L_n^{-\frac{1}{2},-\frac{1}{2}} = \sqrt{2}T_n$. Therefore, if $\alpha = \beta = \frac{1}{2}$, then

$$\sigma_{n,m} = 2\int_{-1}^{1} U_n(t)U_{m-1}(t)\varrho_{\frac{1}{2},\frac{1}{2}}(t)dt = 2\delta_{n+1,m} \quad \Longrightarrow \quad Q_n^{\frac{1}{2},\frac{1}{2}}(\xi) = \frac{2}{z^{n+1}}.$$

Now if $\alpha = \beta = -\frac{1}{2}$. We fix $n \geq 1$ and use the identity $T_{m+1} = \frac{U_{m+1}-U_{m-1}}{2}$ for $m \geq 1$, we infer that for any $m \geq 1$

$$\sigma_{n,m+2} - \sigma_{n,m} = 4\int_{-1}^{1} \sqrt{2}T_nT_{m+1}\varrho_{-\frac{1}{2},-\frac{1}{2}}(t)dt$$

hence for any $j \geq 0$, $\sigma_{n,n+2j} = \sigma_{n,n} = 0$ and $\sigma_{n,n+2j+1} = \sigma_{n,n+1} + 2\sqrt{2} = 2\sqrt{2}$. We deduce that that for any $n \geq 1$

$$Q_n^{-\frac{1}{2},-\frac{1}{2}}(\xi) = \sum_{j=0}^{\infty} \frac{2\sqrt{2}}{z^{n+2j+1}} = 2\sqrt{2}\frac{z}{z^2-1}z^{-n}. \tag{A.3.12}$$

■

The bound of $Q_n^{\alpha,\beta}$ can be sharpened for the values of $\alpha, \beta > 0$ using elementary arguments. We have the following

Lemma A.3.5

Let $\alpha, \beta > 0$. For any $n \geq 1$, we have

$$\sup_{\xi\in\mathcal{E}_s} |Q_n^{\alpha,\beta}(\xi)| = \sup_{z\in\mathcal{U}_s} \left|Q_n^{\alpha,\beta}(\frac{z+z^{-1}}{2})\right| \leq \sqrt{\frac{(\alpha+\beta+1)(\alpha+\beta)}{\alpha\beta}}\frac{s^{-n}}{s-1}. \tag{A.3.13}$$

Proof: Since $U_m(t)\sqrt{1-t^2} \leq 1$ for any $t \in [-1,1]$ and $\varrho_{\alpha,\beta} = \frac{w_{\alpha,\beta}}{W_{\alpha,\beta}}$, then for $m \geq 1$

$$\int_{-1}^{1} U_{m-1}(t)^2\varrho_{\alpha,\beta}(t)dt = \frac{W_{\alpha-1,\beta-1}}{W_{\alpha,\beta}}\int_{-1}^{1} U_{m-1}(t)^2(1-t^2)\varrho_{\alpha-1,\beta-1}(t)dt \leq \frac{W_{\alpha-1,\beta-1}}{W_{\alpha,\beta}} = \frac{(\alpha+\beta)(\alpha+\beta+1)}{4\alpha\beta},$$

where we have used the formula of $W_{\alpha,\beta}$ given in (A.1.2). In view of (A.3.9), this implies $\sigma_{n,m} \leq \sqrt{\frac{(\alpha+\beta)(\alpha+\beta+1)}{\alpha\beta}}$. applying triangular inequality to (A.3.6) implies the result. ■

The previous bound has limitation for values α and β close to 0. We may improve it using finer argument on the growth of some functions associated with Jacobi polynomials.

Lemma A.3.6

Let $\alpha, \beta > -\frac{1}{2}$ and $s > 1$. For any $n \geq 1$, we have

$$\sup_{\xi \in \mathcal{E}_s} |Q_n^{\alpha,\beta}(\xi)| \leq 2C_{\alpha,\beta} c_{\alpha,\beta} \frac{s^{-n}}{s-1}, \tag{A.3.14}$$

with $C_{\alpha,\beta} := \sqrt{e(2+\sqrt{\alpha^2+\beta^2})}$ and $c_{\alpha,\beta} = \frac{(\alpha+\beta+1)(\alpha+\beta+3)}{(2\alpha+1)(2\beta+1)}$. In addition if $\alpha, \beta \leq \frac{1}{2}$, the constant $C_{\alpha,\beta}$ can be replaced by 2.

Proof: The orthonormal Jacobi polynomials with respect to $w_{\alpha,\beta}$ were noted $p_n^{\alpha,\beta}$. We have by (A.1.7) that $L_n^{\alpha,\beta} = \sqrt{W_{\alpha,\beta}} p_n^{\alpha,\beta}$ for any $n \geq 0$. We introduce the function $\varphi_n^{\alpha,\beta}$ defined by

$$\varphi_n^{\alpha,\beta}(t) = (1-t^2)^{\frac{1}{4}} \sqrt{\varrho_{\alpha,\beta}(t)} L_n^{\alpha,\beta}(t) = (1-t^2)^{\frac{1}{4}} \sqrt{w_{\alpha,\beta}(t)} p_n^{\alpha,\beta}(t)$$

For a large class of Jacobi polynomials, it is proven that the function $\varphi_n^{\alpha,\beta}$ is bounded. We do not investigate the best bound possible depending on α and β but use the following bound which is given in [48]; For $\alpha, \beta \geq -\frac{1}{2}$

$$\sup_{t \in [-1,1]} |\varphi_n^{\alpha,\beta}(t)| \leq \sqrt{\frac{2}{\pi}} C_{\alpha,\beta} \quad \text{with} \quad C_{\alpha,\beta} := \sqrt{e(2+\sqrt{\alpha^2+\beta^2})}$$

We introduce the notation $u_m = U_m(t)\sqrt{1-t^2}$. We may bound $\sigma_{n,m}$ using

$$\sigma_{n,m} \leq 2 \int_{-1}^{1} |L_n^{\alpha,\beta}(t)||U_{m-1}(t)| \varrho_{\alpha,\beta}(t) dt = \frac{2}{\sqrt{W_{\alpha,\beta}}} \int_{-1}^{1} |\varphi_n^{\alpha,\beta}||u_{m-1}| w_{\frac{1}{2}\alpha-\frac{3}{4},\frac{1}{2}\beta-\frac{3}{4}} \leq 2\sqrt{\frac{2}{\pi}} C_{\alpha,\beta} \frac{W_{\frac{1}{2}\alpha-\frac{3}{4},\frac{1}{2}\beta-\frac{3}{4}}}{\sqrt{W_{\alpha,\beta}}},$$

where we have used $\frac{1}{2}(\alpha - \frac{3}{2}), \frac{1}{2}(\beta - \frac{3}{2}) > -1$. By Gamma function properties

$$W_{\frac{1}{2}\alpha-\frac{3}{4},\frac{1}{2}\beta-\frac{3}{4}} = 2^{\frac{1}{2}\alpha+\frac{1}{2}\beta-\frac{1}{2}} \frac{\Gamma\left(\frac{1}{2}\alpha+\frac{1}{4}\right)\Gamma\left(\frac{1}{2}\beta+\frac{1}{4}\right)}{\Gamma\left(\frac{1}{2}(\alpha+\beta)+\frac{1}{2}\right)} = c_{\alpha,\beta} 2^{\frac{1}{2}\alpha+\frac{1}{2}\beta+\frac{3}{2}} \frac{\Gamma\left((\frac{1}{2}\alpha+\frac{1}{4})+1\right)\Gamma\left((\frac{1}{2}\beta+\frac{1}{4})+1\right)}{\Gamma\left(\frac{1}{2}(\alpha+\beta)+\frac{1}{2}+2\right)},$$

which is equal to $c_{\alpha,\beta} W_{\frac{1}{2}\alpha+\frac{1}{4},\frac{1}{2}\beta+\frac{1}{4}}$ with $c_{\alpha,\beta} = \frac{(\alpha+\beta+1)(\alpha+\beta+3)}{4(\alpha+\frac{1}{2})(\beta+\frac{1}{2})}$. Using Cauchy Schwartz inequality, we have

$$W_{\frac{1}{2}\alpha+\frac{1}{4},\frac{1}{2}\beta+\frac{1}{4}} = \int_{-1}^{1} \sqrt{w_{\alpha,\beta}(t)}(1-t^2)^{\frac{1}{4}} dt \leq \Big(\int_{-1}^{1} w_{\alpha,\beta}(t) dt\Big)^{\frac{1}{2}} \Big(\int_{-1}^{1} \sqrt{1-t^2} dt\Big)^{\frac{1}{2}} = \sqrt{\frac{\pi}{2}} \sqrt{W_{\alpha,\beta}},$$

hence $\sigma_{n,m} \leq 2C_{\alpha,\beta} c_{\alpha,\beta}$ which implies the first result of the lemma. If in addition $\alpha, \beta \leq \frac{1}{2}$, a sharp bound is known for the function $\varphi_n^{\alpha,\beta}$. Indeed, it is proven in [27] that

$$\sup_{t \in [-1,1]} |\varphi_n^{\alpha,\beta}| \leq 2^{\frac{\alpha+\beta+1}{2}} M_n^{\alpha,\beta} \frac{1}{\sqrt{\pi}} \frac{\Gamma(n+q+1)}{\Gamma(n+1)} N^{-q-\frac{1}{2}} = \sqrt{\frac{2}{\pi}} N^{-q} \left\{ \frac{\Gamma(n+q+1)\Gamma(n+p+q+1)}{\Gamma(n+1)\Gamma(n+p+1)} \right\}^{1/2}.$$

with $p = \min(\alpha, \beta)$, $q = \max(\alpha, \beta)$ and $N = n + \frac{1}{2}(\alpha + \beta + 1)$. Using (A.2.7), we deduce

$$\frac{\Gamma(n+q+1)\Gamma(n+p+q+1)}{\Gamma(n+1)\Gamma(n+p+1)} \leq [(n+q)(n+p+q)]^q \, \frac{1+q+b}{1+a} \leq N^q \frac{\frac{3}{2}+b}{1+a} \leq 2N^q$$

Therefore $\sup_{t\in[-1,1]} |\varphi_n^{\alpha,\beta}| \leq 2\sqrt{\frac{2}{\pi}}$ which implies the second result. ■

For our purpose, the convergence of Jacobi series toward the solution map u in Chapter I and II, the previous bounds are sufficient. Indeed, using very delicate argument, it can shown that the algebraic terms in n can be absorbed using the exponential decay of $|z|^{-n}$ to prove the ℓ^p summability of the sequence of Jacobi coefficient norm in high dimension. However, for numerical experiences, obtaining better bounds yield the construction the Jacobi series that converge to the solution map u rapidly.

Remark A.3.7

The inspection of the various proofs in this section show that the previous result (A.3.13) holds for any probability density function ϱ_0 over $[-1, 1]$ and the corresponding family of orthonormal polynomials $(L_n)_{n\geq 0}$.

We now return to the bounding of $q_n^{\alpha,\beta}$ on the ellipses $\mathcal{E}_s$, $s > 1$ for more general value of $\alpha, \beta > -1$. For this purpose, we may use a crude triangle inequality on the serie (A.3.6) and good bounds on the quantities $\sigma_{m,n}$, which is the method we used in in the previous Lemma to sharpen the bound of $q_n^{0,0}$ with a factor $\frac{1}{\sqrt{2n+1}}$ compared with the bound given [35].

A.4 Growth of quadratic sums associated with Jacobi polynomials

In this section, we give some results related to the growth of certain quantities that we use in chapters 4-5-7. We introduce $\mathcal{F}$ the set of sequences of integer which are infinitely supported, i.e.

$$\mathcal{F} := \Big\{ \nu := (\nu_1, \nu_2, \ldots) : \nu_j \geq 1 \quad \text{and} \quad \#\{j \geq 1 : \nu_j \neq 0\} < \infty \Big\} \tag{A.4.1}$$

and define the order $\leq$ on $\mathcal{F}$ by $\mu \leq \nu$ if and only if $\mu_j \leq \nu_j$ for any $j \geq 1$. Given $\alpha, \beta > -1$, we i introduce the Tensorized Jacobi polynomials indexed in $\mathcal{F}$ by

$$L_\nu^{\alpha,\beta}(y) = \prod_{j\geq 1} L_{\nu_j}^{\alpha,\beta}(y_j), \quad \nu \in \mathcal{F}, \quad y \in U := [-1, 1]^{\mathbb{N}}. \tag{A.4.2}$$

The infinite product is actually finite since $L_0^{\alpha,\beta}$ is constant and equal to 1. We are interested in the growth of the quatity

$$K_{\alpha,\beta}(\Lambda) = \sum_{\nu\in\Lambda} \|L_\nu^{\alpha,\beta}\|^2_{L^\infty(U)} = \sum_{\nu\in\Lambda} \prod_{j\geq 1:\nu_j\neq 0} \|L_{\nu_j}^{\alpha,\beta}\|^2_{L^\infty([-1,1])}, \tag{A.4.3}$$

for lower sets $\Lambda \subset \mathcal{F}$. We recall that Λ is lower if and only if

$$\nu \in \Lambda \quad \text{and} \quad \mu \leq \nu \Rightarrow \mu \in \Lambda. \tag{A.4.4}$$

We first prove good growth bounds for classical cases, namely Legendre polynomials ($\alpha = \beta = 0$) and Tchebychev polynomials ($\alpha = \beta = \pm\frac{1}{2}$), then we provide a growth in the general case, yet very pessimistic. We should note however that for any $\alpha, \beta > -1$ and any $\Lambda \subset \mathcal{F}$, one has

$$K_{\alpha,\beta}(\Lambda) \geq \#(\Lambda). \tag{A.4.5}$$

Indeed, since $\varrho_{\alpha,\beta}$ is a probability measure over [-1,1] then for any $n \geq 1$, $\|L_n^{\alpha,\beta}\|_{L^2([-1,1],d\varrho_{\alpha,\beta})} = 1$ necessarily implies that $\|L_n^{\alpha,\beta}\|_{L^\infty([-1,1])} \geq 1$.

In view of (A.2.2) and (A.2.3), we have

$$K_{0,0}(\Lambda) = \sum_{\nu\in\Lambda} \prod_{j\geq 1} (2\nu_j + 1), \tag{A.4.6}$$

and

$$K_{-\frac{1}{2},-\frac{1}{2}}(\Lambda) = \sum_{\nu\in\Lambda} 2^{\#\,\mathrm{supp}(\nu)}, \qquad K_{\frac{1}{2},\frac{1}{2}}(\Lambda) = \sum_{\nu\in\Lambda} \prod_{j\geq 1} (\nu_j + 1)^2, \tag{A.4.7}$$

We have then the following results

Lemma A.4.1

For any finite lower set $\Lambda \subset \mathcal{F}$, the quantity $K_{0,0}(\Lambda)$ satisfies

$$K_{0,0}(\Lambda) \leq (\#(\Lambda))^2. \tag{A.4.8}$$

Proof: we use induction on $n_\Lambda := \#(\Lambda) \geq 1$. When $n_\Lambda = 1$, then necessarily $\Lambda = \{0_\mathcal{F}\}$ and an equality holds. Let $n \geq 1$ and Λ denote a monotone set with $n_\Lambda = n + 1$. Without loss of generality, we suppose that $\nu_1 \neq 0$ for some $\nu \in \Lambda$. We introduce the indices sets

$$\Lambda_k := \left\{ \hat{\nu} \in \mathcal{F} : (k, \hat{\nu}) \in \Lambda \right\}, \quad k \geq 0. \tag{A.4.9}$$

Here $(k, \hat{\nu})$ denote the multi-index $(k, \hat{\nu}_1, \hat{\nu}_2, \cdots)$. Since Λ is lower and finite, then it is easy to check that the sets Λ_k are finite, lower (when not empty) and satisfy

$$\cdots \subset \Lambda_k \subset \cdots \subset \Lambda_1 \subset \Lambda_0. \tag{A.4.10}$$

Let us also remark that there exists $J \geq 0$ such that $\Lambda_k = \emptyset$ for any $k > J$ and that since $\nu_1 \neq 0$ for some $\nu \in \Lambda$, then $\#(\Lambda_0) \leq n_\Lambda - 1 = n$. Therefore the induction hypothesis applied with the sets Λ_k, implies

$$K(\Lambda) = \sum_{k=0}^{J} (2k+1) K_{0,0}(\Lambda_k) \leq \sum_{k=0}^{J} (2k+1)(\#(\Lambda_k))^2 \,. \tag{A.4.11}$$

Now, by the nestedess of the sets Λ_k, we have

$$k(\#\Lambda_k)^2 \leq \#(\Lambda_k)\#(\Lambda_0) + ... + \#(\Lambda_k)\#(\Lambda_{k-1}), \quad 1 \leq k \leq J. \tag{A.4.12}$$

Therefore

$$K(\Lambda) \leq \sum_{k=0}^{J} (\#(\Lambda_k))^2 + 2 \sum_{k=1}^{J} \sum_{k'=0}^{k-1} \#(\Lambda_k)\#(\Lambda_{k'}) = \Big(\sum_{k=0}^{J} \#(\Lambda_k) \Big)^2. \tag{A.4.13}$$

Using that $\#(\Lambda) = \sum_{k=0}^{J} \#(\Lambda_k)$, we conclude the proof. ■

The previous bound is valid for any lower set and independently of its shape. In addition, the second inequality is sharp, in the sense that equality holds for certain types of lower sets. Indeed, given $\nu \in \mathcal{F}$ supported in $\{1, \cdots, J\}$ and considering the rectanglal block

$$\mathcal{R}_\nu := \{\mu \in \mathcal{F} : \mu \leq \nu\}, \tag{A.4.14}$$

one has

$$K(\mathcal{R}_\nu) = \sum_{\mu \leq \nu} \prod_{1 \leq j \leq J} (2\mu_j + 1) = \prod_{1 \leq j \leq J} \sum_{\mu_j \leq \nu_j} (2\mu_j + 1) = \prod_{1 \leq j \leq J} (\nu_j + 1)^2 = (\#\mathcal{R}_\nu)^2. \tag{A.4.15}$$

Using simply the bound

$$\prod_{j \geq 1} (\nu_j + 1)^2 \leq \prod_{j \geq 1} (\nu_j + 1) \prod_{j \geq 1} (2\nu_j + 1) = \#(\mathcal{R}_\nu) \prod_{j \geq 1} (2\nu_j + 1), \tag{A.4.16}$$

we infer the following

Lemma A.4.2

For any finite lower set $\Lambda \subset \mathcal{F}$, the quantity $K_{\frac{1}{2},\frac{1}{2}}(\Lambda)$ satisfies

$$K_{\frac{1}{2},\frac{1}{2}}(\Lambda) \leq (\#(\Lambda))^3. \tag{A.4.17}$$

This bound is obviously not sharp. We note however that for $\sum_{j=0}^{k} (j+1)^2 \sim k^3$ which show that the exponent 3 can not be improved for all lower sets. We now turn to the case $\alpha = \beta = -\frac{1}{2}$ which corresponds to Tchebychev polynomials of first kind. We prove an intermediate proposition then provide the growth.

Proposition A.4.3

For any real positive numbers $a_0 \geq a_1 \geq ... \geq a_k$ and any $\gamma \geq \frac{\ln 3}{\ln 2}$, one has

$$a_0^\alpha + 2(a_1^\gamma + \ldots + a_k^\gamma) \leq (a_0 + \ldots + a_k)^\gamma. \tag{A.4.18}$$

Proof: We use induction on k. For $k = 0$, an equality holds in (A.4.18). For $k = 1$, since the function $x \mapsto (x + a_1)^\alpha - x^\alpha$ is increasing in $[a_1, +\infty[$ then its value at a_0 is greater than its value at a_1, that is

$$2a_1^\gamma \leq (2^\gamma - 1)a_1^\gamma \leq (a_0 + a_1)^\gamma - a_0^\gamma \tag{A.4.19}$$

where we have used $2^\gamma > 3$. Now let $k \geq 1$ and $a_0 \geq a_1 \geq ... \geq a_{k+1}$ be real positive numbers. By the induction hypothesis at steps 1 and k, we infer

$$\begin{aligned}(a_0 + ... + a_{k+1})^\gamma &= \Big((a_0 + ... + a_k) + a_{k+1}\Big)^\gamma \\ &\geq (a_0 + ... + a_k)^\gamma + 2a_{k+1}^\gamma \\ &\geq a_0^\gamma + 2(a_1^\gamma ... + a_k^\gamma) + 2a_{k+1}^\gamma \\ &= a_0^\gamma + 2(a_1^\gamma ... + a_{k+1}^\gamma).\end{aligned} \tag{A.4.20}$$

The proof is then complete. ■

Lemma A.4.4

For any lower set $\Lambda \subset \mathcal{F}$, the quantity $K_{-\frac{1}{2},-\frac{1}{2}}(\Lambda)$ satisfies

$$K_{-\frac{1}{2},-\frac{1}{2}}(\Lambda) \leq (\#(\Lambda))^{\frac{\log 3}{\log 2}}. \tag{A.4.21}$$

Proof: We use induction on $n_\Lambda := \#(\Lambda)$. When $n_\Lambda = 1$, then necessarily $\Lambda = \{0_\mathcal{F}\}$ and an equality holds. Let $n \geq 1$ and Λ denote a lower set with $n_\Lambda = n + 1$. Without loss of generality, we suppose that $\nu_1 \neq 0$ for some $\nu \in \Lambda$. Defining $J \geq 0$ and the sets Λ_k as in the proof of Lemma A.4.1 and using the induction hypothesis with these sets, we obtain

$$K_{-\frac{1}{2},-\frac{1}{2}}(\Lambda) = \sum_{k=0}^{J} \gamma(k) K_{-\frac{1}{2},-\frac{1}{2}}(\Lambda_k) \leq \sum_{k=0}^{J} \gamma(k)(\#(\Lambda_k))^{\frac{\log 3}{\log 2}}, \tag{A.4.22}$$

where γ is defined by $\gamma(0) = 1$ and $\gamma(k) = 2$ for $k \geq 1$. Using (A.4.18), we infer

$$K_{-\frac{1}{2},-\frac{1}{2}}(\Lambda) \leq (\#(\Lambda_0))^{\frac{\log 3}{\log 2}} + 2\sum_{k=1}^{J}(\#(\Lambda_k))^{\frac{\log 3}{\log 2}} \leq \Big(\#(\Lambda_0) + \#(\Lambda_1) + \cdots + \#(\Lambda_J)\Big)^{\frac{\log 3}{\log 2}} = (\#(\Lambda))^{\frac{\log 3}{\log 2}}. \tag{A.4.23}$$

The proof is then complete. ■

The previous bound (A.4.21) can also be sharp for certain type of lower sets. For instance if ν is the multi-index such that $\nu_1 = \ldots = \nu_J = 1$ and $\nu_j = 0$ for $j > J$, then

$$K_{-\frac{1}{2},-\frac{1}{2}}(\mathcal{B}_\nu) = \sum_{\mu\le\nu} 2^{\#\mathrm{supp}(\mu)} = \sum_{\mu\le\nu} 2^{\mu_1+\cdots+\mu_J} = \prod_{j=1}^{J}(1+2) = 3^J = (2^J)^\beta = (\#\mathcal{B}_\nu)^\beta. \tag{A.4.24}$$

For more general values of α and β, we have shown in (A.2.4), that

$$\|L_n^{\alpha,\beta}\|_{L^\infty} \le C_{\alpha,\beta}(n+1)^{\max(q+\frac{1}{2},0)}, \quad q = \max(\alpha,\beta), \quad n \ge 1. \tag{A.4.25}$$

where $C_{\alpha,\beta}$ depending only on α and β. Since $\prod_{j\ge1}(\nu_j+1)^{\max(q+\frac{1}{2},0)} = (\#(\mathcal{B}_\nu))^{\max(q+\frac{1}{2},0)}$, then a rough bound on $K_{\alpha,\beta}(\Lambda)$ for lower sets Λ is given by

$$K_{\alpha,\beta}(\Lambda) \le (\#(\Lambda))^{\max(2q+1,0)} \sum_{\nu\in\Lambda} C_{\alpha,\beta}^{2\#(\mathrm{supp}\,\nu)} \le (\#(\Lambda))^{\max(2q+1,0)+\gamma}, \quad \gamma = \frac{\log(C_{\alpha,\beta}^2+1)}{\log 2} \tag{A.4.26}$$

where we have used the same argument as in the previous lemma exploiting that for any real positive numbers $a_0 \ge a_1 \ge ... \ge a_k$ one has

$$a_0^\gamma + C_{\alpha,\beta}(a_1^\gamma + \ldots + a_k^\gamma) \le (a_0 + \ldots + a_k)^\gamma, \tag{A.4.27}$$

which can be proved by induction as in Proposition A.4.3

Bibliography

[1] R. Andreev, M. Bieri, and C. Schwab. Sparse tensor discretization of elliptic spdes. *SIAM J. Sci. Comput*, 31:4281–4304, 2006.

[2] I. Babuška and P. Chatzipantelidis. On solving elliptic stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 191(37-38):4093 – 4122, 2002.

[3] I. Babuška, M. K. Deb, and J.T. Oden. Solution of stochastic partial differential equations using galerkin finite element techniques. *Computer Methods in Applied Mechanics and Engineering*, 190(48):6359–6372, 2001.

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

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

[6] I. Babuška, R. Tempone, and Zouraris G. E. Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Computer Methods in Applied Mechanics and Engineering*, 194(12-16):1251–1294, 2005.

[7] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. On the optimal polynomial approximation of stochastic pdes by galerkin and collocation methods. *Math. Mod. Methods Appl. Sci.*, 2011.

[8] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. Stochastic spectral galerkin and collocation methods for pdes with random coefficients: A numerical comparison. In *Spectral and High Order Methods for Partial Differential Equations*, volume 76, pages 43–62. Springer Berlin Heidelberg, 2011.

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

[10] N. Batir. Inequalities for the gamma function. *Archiv der Mathematik.*, 91-6:554–563, 2008.

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

[12] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM Journal on Mathematical Analysis*, 43(3):1457–1472, 2011.

[13] P. Binev, W. Dahmen, and R. DeVore. Adaptive finite element methods with convergence rates. *Numerische Mathematik*, 97(2):219–268, 2004.

[14] L. Bos and N. Levenberg. On the calculation of approximate fekete points: the univariate case. *Electronic Transactions on Numerical Analysis*, 30:377–397, 2008.

[15] S. Brenner and L.R Scott. *The mathematical theory of Finite Elements*. Springer, second edition, 2008.

[16] A. Buffa, Y. Maday, A.T. Patera, C. Prudhomme, and G. Turinici. A priori convergence of the greedy algorithm for the parameterized reduced basis. *M2AN*, 46-3:595–603, 2012.

[17] J.P. Calvi and L. Bialas-Ciez. Pseudo leja sequence. *Ann. Mat. Pura Appl*, 191:53–75, 2012.

[18] J.P. Calvi and V.M. Phung. On the lebesgue constant of leja sequences for the unit disk and its applications to multivariate interpolation. *Journal of Approximation Theory*, 163-5:608–622, 2011.

[19] J.P. Calvi and V.M. Phung. Lagrange interpolation at real projections of leja sequences for the unit disk. *Proceedings of the American Mathematical Society*, 2012.

[20] T. S. Chihara. *An introduction to orthogonal polynomials*. Gordon and Breach, 1978.

[21] A. Chkifa. On the lebesgue constant of leja sequences for the complex unit disk and of their real projection. *Journal of Approximation Theory*, 166:176–200, 2013.

[22] A. Chkifa, A. Cohen, R. DeVore, and C. Schwab. Sparse adaptive taylor approximation algorithms for parametric and stochastic elliptic pdes. *M2AN*, 2012.

[23] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile, and R. Tempone. Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic pdes. *submitted*, 1:1–2, 2014.

[24] A. Chkifa, A. Cohen, P.Y. Passaggia, and J. Peter. A comparative study between kriging and adaptive sparse tensor-product methods for high dimensional approximation and optimisation problems in aerodynamics design. *ESAIM proceedings*.

[25] A. Chkifa, A. Cohen, and C. Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric pdes. *Pure. Applied. Math*, 2013.

[26] A. Chkifa, A. Cohen, and C. Schwab. High-dimensional adaptive sparse polynomial interpolation and applications to parametric pdes. *Foundations of Computational Mathematics*, pages 1–33, 2013.

[27] Y. Chow, L. Gatheschi, and R. Wong. A bernstein type inequality for the jacobi polynomial. *Proc Amer Math Soc*, 121:703–709, 1994.

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

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

[30] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods for elliptic operator equations convergence rates. *Math. Comp.*, 70:27–75, 2001.

[31] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods for elliptic operator equations ii beyond the elliptic case. *Found. Comput. Math.*, 2:203–345, 2002.

[32] A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least square approximations. *Found. Comp. Math*, 152:621–647, 2004.

[33] A. Cohen, R. DeVore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic pde's. *Analysis and Application*, 2010.

[34] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best n-term galerkin approximations for a class of elliptic spdes. *J. FoCM*, 2010.

[35] P. J. Davis. *Interpolation and Approximation*. Blaisdell Publishing Company, 1963.

[36] C. De Boor and A. Ron. Computational aspects of polynomial interpolation in several variables. *Mathematics of Computation*, 58:705–727, 1992.

[37] S. De Marchi. On leja sequences: some results and applications. *Applied Mathematics and Computation*, 152:621–647, 2004.

[38] V. De Silva and L. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

[39] M. K. Deb. *Solution of stochastic partial differential equations (SPDEs) using Galerkin method: theory and applications*. Ph.D. Dissertation, The University of Texas, Austin, 2000.

[40] R. DeVore. *Nonlinear Approximation*, volume 7. Acta Numerica, 1998.

[41] R. DeVore and G.G. Lorentz. *Constructive Approximation*. Springer, 1993.

[42] R. DeVore, G. Petrova, and P. Wojtaszczyk. Greedy algorithms for reduced bases in banach spaces. *Constructive Approximation*, 37(3):455–466, 2013.

[43] R. DeVore, Howard R., and C. Micchelli. Optimal nonlinear approximation. *Manuscripta Math*, 63:469–478, 1989.

[44] S. Dineen. *Complex Analysis on Infinite Dimensional Spaces*. Springer Monographs in Mathematics, Springer Verlag, Berlin, 1999.

[45] A. Doostan, R. Ghanem, and J. Red-Horse. Stochastic model reduction for chaos representations. *Computer Methods in Applied Mechanics and Engineering*, 196(37-40):3951 – 3966, 2007.

[46] W. Dorfler. A convergent adaptive algorithm for poisson's equation. *SIAM Journal on Numerical Analysis*, 33(3):1106–1124, 1996.

[47] V.K. Dzjadyk and V.V. Ivanov. On asymptotics and estimates for the uniform norms of the lagrange interpolation polynomials corresponding to the chebyshev nodal points. *Analysis Mathematica*, 9-11:85–97, 1983.

[48] T. Erdélyi, A. P. Magnus, and P. Nevai. Generalized jacobi weights, christoffel functions, and jacobi polynomials. *SIAM J. Math. Anal*, 25:602–614, 1994.

[49] T. Gantumur, H. Harbrecht, and R. Stevenson. An optimal adaptive wavelet methods without coarsening of the iterands. *Mathematics of computation*, 76:615–629, 2007.

[50] T. Gerster and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71-1, 2003.

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

[52] R. Ghanem and P. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Dover, New York, second edition, 2002.

[53] C.J Gittelson. Spectral techniques for stochastic finite elements. *Math. Comp*, 2012.

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

[55] T.H. Gronwall. A sequence of polynomials connected with the n-th roots of unity. *Bull. Amer. Math. Soc.*, 27:275–279, 1921.

[56] V.H. Hoang and C. Schwab. Sparse tensor galerkin discretizations for parametric and random parabolic pdes i: Analytic regularity and gpc-approximation. *Seminar for Applied Mathematics, ETH Zurich*, 2010-11.

[57] V.H. Hoang and C. Schwab. Analytic regularity and gpc approximation for parametric and random 2nd order hyperbolic pdes. *Analysis and Applications*, 2011.

[58] J. Kuntzman. *Méthodes numériques - Interpolation, dérivées*. Dunod, Paris, 1959.

[59] O.P. Le Maître and O.M. Knio. *Spectral Methods for Uncertainty Quantification*. Scientific Computation. Springer Netherlands, 2010.

[60] G. Lorentz and R. Lorentz. Solvability problems of bivariate interpolation I. *Constructive Approximation*, 2:155–169, 1986.

[61] Ledoux M. and Talagrand M. *Probability in Banach spaces*. Springer Verlag, Berlin, 1991.

[62] Y. Maday, N.C. Nguyen, A.T. Patera, and G.S.H. Pau. A general multipurpose interpolation procedure: the magic points. *Commun. Pure Appl. Anal*, 8:383–404, 2009.

[63] Y. Maday, A.T. Patera, and G. Turinici. Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations. *Comptes Rendus Mathematique*, 335(3):289–294, 2002.

[64] Y. Maday, A.T. Patera, and G. Turinici. A priori convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations. *Journal of Scientific Computing*, 17(1-4):437–446, 2002.

[65] H. G. Matthies and A. Keese. Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 194(12-16):1295 – 1331, 2005.

[66] G. Migliorati, F. Nobile, E. von Schwerin, and R. Tempone. Analysis of the discrete l^2 projection on polynomial spaces with random evaluations. *MOX Report 46/2011*, 2011.

[67] P Morin, R. H. Nochetto, and K. G. Siebert. Data oscillation and convergence of adaptive fem. *SIAM Journal on Numerical Analysis*, 38:466–488, 1999.

[68] F. Nobile and R. Tempone. Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients. *International Journal for Numerical Methods in Engineering*, 80(6-7):979–1006, 2009.

[69] 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.

[70] 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.

[71] E. Novak and K. Ritter. High dimensional integration of smooth functions over cubes. *Numerische Mathematik*, 75(1):79–97, 1996.

[72] T. Runst and W. Sickel. *Sobolev spaces of fractional order, Nemytskij operators, and nonlinear partial differential equations.* De Gruyter series in nonlinear analysis and applications, De Gruyter, Berlin, 1996.

[73] C. Schwab and R.S. Stevenson. Space-time adaptive wavelet methods for parabolic evolution equations. *Math. Comp.*, 78:1293–1318, 2009.

[74] C. Schwab and R.A. Todor. Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA Journal of Numerical Analysis*, 27:232–261, 2007.

[75] J.S. Smith. Lebesgue constants in polynomial interpolation. *Annales Mathematicae et Informaticae*, 33:109–123, 2006.

[76] S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Doklady Akademii Nauk SSSR 4*, pages 240–243, 1963.

[77] A. Sommariva and M. Vianello. Approximate fekete points for weighted polynomial interpoation. *Electronic Transactions on Numerical Analysis*, 37:1–22, 2010.

[78] R. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.

[79] G. Szego. *Orthogonal Polynomials.* American Mathematical Society Colloquium Publication, forth edition, 1975.

[80] V. Temlyakov. Nonlinear kolmogorov width. *Math. Notes*, 63-6:785–795, 1998.

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

[82] D. Xiu and G. E. Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191(43):4927–4948, 2002.