Approximations de données en grande dimension
Notes du cours de M2

Albert Cohen

L'objectif du cours est l'étude du problème général suivant : *reconstruire approximativement* une fonction u inconnue, définie sur un domaine $D \subset \mathbb{R}^d$, à partir de *données ponctuelles* en des points $x^1, \ldots, x^m \in D$. Ce problème très général devient souvent difficile en grande dimension $d >> 1$. Le cours se divise en 5 chapitres.

1. Introduction au cours.
2. Notions de théorie de l'approximation linéaire et non-linéaire.
3. Reconstruction à partir de données ponctuelles et échantillonage aléatoire.
4. Espaces linéaires optimaux et bases réduites.
5. Résultats d'approximation en dimension infinie pour les EDP paramétriques.

Les chapitres 1 et 2 traitent de notions classiques. Les chapitres 3, 4 et 5 abordent des résultats de recherches récents (ultérieurs à 2010). Le cours aborde à la fois des aspects de probabilités (chapitre 3), d'analyse numérique, et de modélisation par les équations aux dérivées partielles (chapitre 5). Quelques connaissances de bases en probabilités et analyse de niveau L3 et M1 sont supposées comme acquises.

Références bibliographiques pour approfondir :

• [1] R. DeVore, "Nonlinear approximation", Acta Numerica, 1998.
• [2] A. Cohen and G. Migliorati, "Optimal weighted least-squares methods", SMAI Journal of Computational Mathematics, 2017.
• [3] A. Cohen and R. DeVore, "High dimensional approximation of parametric PDEs", Acta Numerica, 2015.
• [4] J. Tropp, "An introduction to matrix concentration inequalities", Foundations and trends in machine learning, 2015.

1 Introduction au cours

On reconstruit donc une fonction $\tilde{u} \neq u$ à partir de données ponctuelles et on étudie l'erreur $\|u - \tilde{u}\|$ dans une certaine norme $\|\cdot\| = \|\cdot\|_V$, associée à un espace de Banach V, par exemple l'erreur L^∞ ou $L^2(D, \mu)$ où μ est une mesure définie sur D. On s'intéresse en particulier à la vitesse de convergence vers 0 de l'erreur quand $m \to \infty$. Ce problème est très général, nous allons donner des cadres mathématiques correspondant à des situations plus précises.

1.1 Exemples de problèmes d'estimation/reconstruction

Les techniques et outils d'analyse qu'on va discuter trouvent des applications dans les trois cadres distinct suivants :

1. *Régression :* on observe m réalisations (x^i, y^i) indépendantes d'une variable aléatoire $z = (x, y)$ de loi jointe ρ inconnue, avec $x \in D$ et $y \in \mathbb{R}$. La variable de sortie y ne dépend pas de façon déterministe de x, ce qui signifie par exemple qu'il manque des paramètres dans le vecteur x pour prédire complètement y (exemple : le réchauffement climatique en un point du globe ne dépend pas uniquement des émmisions de CO2 en ce point). En d'autres termes, il n'existe pas de fonction u telle que $y = u(x)$. On cherche néammoins à recontruire une fonction u qui "explique" au mieux y en fonction de x au sens où elle rend $u(x)$ proche de y dans un sens probabiliste, par exemple au sens du risque quadratique moyen

$$R(v) := \mathbb{E}(|v(x) - y|^2). \tag{1.1}$$

Un calcul élémentaire (**exercice**) montre que la *fonction de régression* définie par l'espérance conditionelle $u(x) := \mathbb{E}(y|x)$ minimise R parmis toutes les fonctions v. Indication : si z est une variable aléatoire la constante $c \in \mathbb{R}$ qui minimise $\mathbb{E}(|z - c|^2)$ est $c = \mathbb{E}(z)$, et pour toute constante $d \in \mathbb{R}$ on a le théorème de Pythagore $\mathbb{E}(|z - d|^2) = \mathbb{E}(|z - c|^2) + |c - d|^2 = \mathrm{var}(z) + |c - d|^2$. Si la mesure ρ a la forme $d\rho(x, y) = p(x, y)dxdy$ où p est une fonction de densité continue et intégrable, alors cette quantité s'écrit

$$u(x) = \int_{\mathbb{R}} y p_x(y) dy, \tag{1.2}$$

où $p_x(y) := [\int_{\mathbb{R}} p(x, y)dy]^{-1} p(x, y)$ est la densité de probabilité de y conditionellement à x. Comme la loi ρ, cette fonction nous est inconnue, et pour toute fonction $\tilde{u}$, le théorème de Pythagore donne (**exercice**)

$$R(\tilde{u}) = R(u) + \mathbb{E}(|u(x) - \tilde{u}(x)|^2) = R(u) + \|u - \tilde{u}\|^2_{L^2(D,\mu)}, \tag{1.3}$$

où μ est la mesure de probabilité marginale de la variable x, c'est à dire

$$d\mu(x) = P(x)dx, \quad P(x) = \int_{\mathbb{R}} p(x, y)dy. \tag{1.4}$$

Noter que ρ est le produit des mesures marginale et conditionelle : $d\rho(x, y) = p_x(y)d\mu(x)$. Il est donc naturel d'essayer de controler l'erreur entre u et $\tilde{u}$ en norme $L^2(D, \mu)$. Noter qu'on peut penser à ce problème comme à l'estimation de u à partir d'évaluation ponctuelles bruitées, en écrivant $y^i = u(x^i) + \eta^i$ où η^i est la réalisation d'une variable η centrée ($\mathbb{E}(\eta|x) = 0$ pour chaque x) qu'on peut interprêter comme un bruit additif.

2. *Planification d'expérience physique ou numérique :* On s'intéresse à approcher une fonction u inconnue de d variables représentées par $x = (x_1, \ldots, x_d) \in D$, que l'on peut évaluer (à une certaine précision près) en des points $x^i \in D$ que l'on choisit. Typiquement, l'évaluation de cette fonction est le résultat d'un calcul numérique ou d'une expérience physique dépendant de paramètres d'entrée décrits par les variables $x_1, \ldots, x_d$, et elle est donc coûteuse. Exemple : les variables x_j décrivent les diverse concentration d'un mélange constituant un carburant et $u(x)$ est la

consommation moyenne d'un moteur quand on utilise ce mélange. On cherche donc à obtenir à partir d'un petit nombre d'évaluation de u un *modèle simplifié* $\tilde{u}$ dont le calcul en tout point sera rapide. On pourra chercher à contrôler l'erreur uniformément sur $x \in D$ ce qui revient à utiliser la norme $L^\infty(D)$, ou en un sens moyen, par exemple en norme $L^2(D,\mu)$ pour une certaine mesure μ, ce qui est en particulier pertinent lorsque les paramètres composant le vecteur x sont aléatoires et distribués suivant une loi de probabilité de mesure μ. Notons (**exercice**) qu'on a toujours la majoration

$$\|u - \tilde{u}\|_{L^2(D,\mu)} \leq \|u - \tilde{u}\|_{L^\infty(D)},$$

ce qui veut simplement dire que l'erreur moyenne est contrôlée par l'erreur uniforme. Contrairement au cas de la régression, il y a ici un phénomène déterministe décrit par la fonction u dont les observations peuvent être supposées non-bruitées. Une autre différence fondamentale avec le cadre de la régression est que nous avons ici le *choix* des points d'évaluation x^i (apprentissage "actif") et on peut à chercher à comprendre si il existe un choix optimal dans un sens à préciser.

Un cadre mathématique important est celui où $u(x)$ est la solution d'une équation aux dérivées partielles (EDP) dépendant de paramètres $x = (x_1, \ldots, x_d) \in D \subset \mathbb{R}^d$. Un exemple simple est l'équation elliptique de diffusion stationaire

$$-\mathrm{div}(a\nabla u)(z) = f(z), \quad z \in \Omega \tag{1.5}$$

posée sur un domaine physique $\Omega \subset \mathbb{R}^2$ ou $\mathbb{R}^3$ (avec des conditions aux limites prescrite), et où les paramètres x_j entrent en jeu pour décrire la fonction de diffusion a, par exemple sous la forme affine

$$a = a(x) = \overline{a} + \sum_{j=1}^{d} x_j \psi_j, \tag{1.6}$$

où $\overline{a}$ et ψ_j sont des fonctions fixées définies sur Ω. Ainsi pour chaque $x \in D$, l'évaluation de la solution $u(x)$ correspondante nécessite de résoudre l'EDP, typiquement par une méthode numérique (élements finis par exemple) qui est coûteuse si on souhaite une grande précision. Notons que pour chaque x donné, la sortie $u(x)$ n'est pas un nombre réel mais une fonction toute entiere de la variable spatiale $z \in \Omega$, appartenant à un espace de dimension infinie H, ou son approximation sur une grille de discrétisation spatiale. Par exemple, dans le cas de l'EDP de diffusion ci-dessus, il est naturel d'utiliser l'espace de Sobolev $H = H^1(\Omega)$. On cherche donc à fabriquer une fonction $x \mapsto \tilde{u}(x)$ simple à calculer à partir de quelques solutions particulières $u(x^i)$. Les norme L^∞ et L^2 s'entendent donc ici au sens $L^\infty(D,H)$ et $L^2(D,H,d\mu)$, c'est à dire

$$\|v\|_{L^\infty} = \sup_{x\in D} \|v(x)\|_H \quad \text{et} \quad \|v\|_{L^2} = \Big(\int_D \|v(x)\|_H^2 d\mu\Big)^{1/2}. \tag{1.7}$$

Il est aussi fréquent qu'on ne cherche pas à approcher toute la solution $u(x) \in H$ mais simplement une quantité scalaire $q(x) = l(u(x)) \in \mathbb{R}$ dependant de celle-ci,

où l est typiquement une forme linéaire : par exemple sa valeur en un point de Ω, ou son intégrale sur Ω ou sur une partie de Ω. On revient ainsi aux normes L^∞ et L^2 classique des fonctions à valeurs dans $\mathbb{R}$, mais il faut noter que l'évaluation de la quantité $q(x) \in \mathbb{R}$ passe a-priori par le calcul exact ou approché de la fonction $u(x) \in H$.

3. *Assimilation de données et problèmes inverses :* On cherche à estimer une fonction u représentant un phénomène physique qu'on observe partiellement. Par exemple la quantité $u(x)$ peut représenter la concentration d'un polluant au point x d'une ville représentée par le domaine D, ou la distribution de température, ou le niveau sonore. Les données sur u peuvent être des évaluations ponctuelles capturées en des points x^i du domaine physique D, où plus généralement des quantités

$$l_i(u), \quad i = 1, \ldots, m, \tag{1.8}$$

où les l_i sont des formes linéaires qui modélisent le dispositif de mesure. Il peut par exemple s'agir de moyennes locales autours de points x^i, c'est à dire de la forme $\int u(x)\varphi(x - x_i)dx$ où φ est une fonction concentrée autour de l'origine et telle que $\int \varphi = 1$, ou d'intégrales le long de lignes dans le cas de mesures tomographiques (transformée de Radon) de la densité de tissus d'un individu vivants. Bien entendu la variable x peut aussi intégrer une variable de temps lorsque le phénomène physique est évolutif. Comme dans le cadre précèdent, on est dans un cadre déterministe : il y a une vraie fonction u qui décrit la quantité physique en fonction de x. Il est cependant plus fréquent de se trouver en présence de données bruitées à cause des erreurs de mesures inhérentes aux capteurs. Par ailleurs, on cherche à à combiner les données avec la connaissance physique du phénomène : par exemple on sait que la distribution de température est solution d'une équation de la chaleur (même si il nous manque typiquement des paramètres - conditions initiales, conditions aux limites, conductivité thermique - pour la résoudre. Dans toutes ces applications, on peut aussi se poser la question du choix optimal des mesures à effectuer, c'est à dire du positionnement optimal des capteurs.

Dans tous les cas, les données sont insuffisantes pour caractériser la fonction u : une infinité de fonctions peuvent avoir les mêmes valeurs en un nombre fini de point. Afin d'espérer reconstruire u avec une certaine précision, il est nécessaire de limiter son caractère arbitraire en introduisant de l'information *a-priori* qui peut s'exprimer de diverses manières. On peut par exemple supposer que u appartient à une classe restreinte $\mathcal{K} \subset V$ reflettant certaines propriétés : u est positive, u a une certaine régularité. On peut aussi chercher exploiter un modèle physique (EDP) qui décrit la fonction u, dans les cadres 2 et 3 décrits ci-dessus, ce qui correspond aussi à rechercher u dans une classe $\mathcal{K}$ représentant l'ensemble des solutions de ce modèle.

1.2 Approximation et reconstruction

Dans de nombreux cas, la connaissance de la classe $\mathcal{K}$ n'est pas suffisante pour reconstruire exactement u (par exemple il existe une infinité de fonction positives de classe $\mathcal{C}^2$ qui ont les mêmes valeurs en un nombre finis de points). On est ainsi amené à rechercher une

approximation dans un modèle *réduit ou simplifié* $V_n \subset V$. L'ensemble V_n est typiquement un espace de dimension finie n (un exemple simple : les polynômes de degré $n-1$ quand on travaille en dimension 1), ou plus généralement un ensemble de fonctions qui peuvent être décrites par un nombre fini n de paramètres. On cherchera donc à reconstruire $\tilde{u} \in V_n$, et on sera amené à comparer $\|u - \tilde{u}\|_V$ avec l'erreur de meilleure approximation

$$e_n(u) = e_n(u)_V := \min_{v \in V_n} \|u - v\|_V. \tag{1.9}$$

On montrer facilement (**exercice**) que ce min est atteint dans le cas d'un sous-espace de dimension finie n. En pratique on considère souvent une suite d'espaces d'approximation $(V_n)_{n \geq 0}$ parfois emboitée au sens où $V_n \subset V_{n+1}$ et on s'attend ainsi à ce que $e_n(u)$ tend vers 0 quand $n \to +\infty$ On peut penser à $e_n(u)$ comme une *erreur de modèle* puisqu'elle prend en compte le fait que la vraie fonction u peut ne pas appartenir à l'espace V_n.

Remarque 1.1 *Pour une classe de fonction $\mathcal{K} \subset V$ donnée, on pourra considérer l'erreur d'approximation maximale sur $\mathcal{K}$, c'est à dire*

$$e_n(\mathcal{K}) = e_n(\mathcal{K})_V := \max_{u \in \mathcal{K}} e_n(u)_V. \tag{1.10}$$

*et on peut montrer (***exercice***) que ce max est atteint dans le cas où $\mathcal{K}$ est un compact de V.*

On a évidemment $\|u - \tilde{u}\|_V \geq e_n(u)$. Nous dirons qu'une méthode de reconstruction a une précision quasi-optimale si pour toute fonction u, sa recontruction $\tilde{u}$ vérifie

$$\|u - \tilde{u}\|_V \leq C_0 e_n(u)_V \tag{1.11}$$

où $C_0 \geq 1$ est une constante fixée. Une telle propriété sous entend que $\tilde{u} = u$ lorsque $u \in V_n$ c'est à dire que toute fonction de V_n est entièrement caractérisée par ses m valeurs aux points $x^1, \ldots, x^m$, ce qui n'est possible que si $m \geq n$. On dira que le budget d'échantillonage est quasi-optimal si $m \leq C_1 n$ où C_1 est une constante fixée. On verra qu'il est difficile d'avoir des méthodes de reconstruction combinant précision quasi-optimale et budget quasi-optimal mais qu'on peut s'en approcher.

Arrêtons nous sur la procédure de reconstruction la plus élémentaire qui consiste à *interpoler* les données. On prend dans ce cas $m = n$ points de mesures $x^1, \ldots, x^n \in D$ dont on suppose qu'ils ont la propriété suivante : l'application

$$L : v \mapsto (v(x^1), \ldots, v(x^n)) \tag{1.12}$$

est un isomorphisme de V_n vers $\mathbb{R}^n$. Cela exige en particulier que les x^j soient distincts mais ce n'est pas toujours suffisant (**exercice**). On travaille ici dans l'espace $V = \mathcal{C}(D)$ de fonctions continues, muni de la norme L^∞. On peut alors définir l'interpolation de données $(y^1, \ldots, y^n)$ associées aux points $(x^1, \ldots, x^n)$ comme l'unique élément $v \in V_n$ tel que

$$v(x^i) = y^i, \quad i = 1, \ldots, n. \tag{1.13}$$

On notera $\tilde{u} = R_n(y^1, \ldots, y^n)$ ce qui définit ainsi un opérateur lineaire de reconstruction $R_n : \mathbb{R}^n \to V_n$. Lorsque les données sont les évaluations non bruitées d'une fonction $u \in V$, c'est à dire $y^i = u(x^i)$ on pose alors $R_n(y^1, \ldots, y^n) = I_n u$, l'unique élément de V_n tel que

$$I_n u(x^i) = u(x^i), \quad i = 1, \ldots, n. \tag{1.14}$$

L'opérateur I_n est linéaire, et vérifie $I_n v = v$ pour tout $v \in V_n$, on peut ainsi le voir comme un projecteur (non-orthogonal).

Il est utile d'exprimer les opérateurs I_n et R_n au moyen des fonctions de base de Lagrange $\{\ell_1, \ldots, \ell_n\}$ de V_n qui sont définies par les relations $\ell_i(x^j) = \delta_{i,j}$. On a alors

$$R_n(y^1, \ldots, y^n) = \sum_{i=1}^n y^i \ell_i \quad \text{et} \quad I_n v = \sum_{i=1}^n v(x^i)\ell_i. \tag{1.15}$$

L'interpolation utilise un budget optimal de mesure puisque $m = n$.

Etudions sa précision en norme L^∞. Celle-ci est intimement liée aux normes des opérateurs R_n et I_n. On voit facilement (**exercice**) que

$$\Lambda_n := \|R_n\|_{\ell^\infty \to L^\infty} = \|I_n\|_{L^\infty \to L^\infty} = \sup_{x \in D} \sum_{i=1}^n |\ell_i(x)| = \Big\| \sum_{i=1}^n |\ell_i| \Big\|_{L^\infty(D)}. \tag{1.16}$$

La quantité Λ_n, qui est toujours supérieure à 1, est appellée *constante de Lebesgue*, elle dépend de l'espace V_n et du choix des points d'interpolation $x^1, \ldots, x^n$. Elle quantifie la stabilité du procédé d'interpolation. Son importance pour l'étude de précision vient de la remarque suivante : si $u \in V$, on a pour tout $v \in V_n$

$$\|u - I_n u\|_{L^\infty} \leq \|u - v\|_{L^\infty} + \|v - I_n u\|_{L^\infty} = \|u - v\|_{L^\infty} + \|I_n(v - u)\|_{L^\infty} \leq (1 + \Lambda_n)\|u - v\|_{L^\infty}. \tag{1.17}$$

Comme v est arbitraire, il vient

$$\|u - I_n u\|_{L^\infty} \leq (1 + \Lambda_n) e_n(u)_V. \tag{1.18}$$

Ainsi, lorsqu'on reconstruit u à partir de données ponctuelles non-bruitées, l'approximation obtenue $\tilde{u} = I_n u$ a une précision sous-optimale car l'amplification par la constante de Lebesgue peut augmenter avec n. Cela signifie que même si l'erreur de meilleure approximation $e_n(u)$ décroit rapidement vers 0 lorsque $n \to \infty$, ceci n'est en rien garanti pour l'erreur d'interpolation qui peut même dans certain cas exploser lorsque Λ_n croît trop rapidement.

Illustrons autour de quelques exemples simples le "conflit" entre la décroissance avec n de l'erreur de d'approximation $e_n(u)_V$ et la croissance possible de la constante de Lebesgue Λ_n :

1. L'erreur d'approximation $e_n(u)_V$ ne dépend que de la fonction u et des espaces V_n considérés, elle ne depend pas des points x^i. Sa vitesse de décroissance est souvent reliée à la régularité de u dans les limites des possibilités d'approximation des espaces V_n. Par exemple si D est un intervalle fermé borné, et V_n les polynômes

de degré $n-1$, il est bien connu que l'erreur d'approximation pour la norme L^∞ décroit comme n^{-k} si u est de classe $\mathcal{C}^k(D)$, avec une estimation de la forme

$$e_n(u)_{L^\infty} \leq Cn^{-k}\|u^{(k)}\|_{L^\infty}, \tag{1.19}$$

où C est une constante qui dépend de D et k. Ceci nous montre en particulier que si $\mathcal{K}$ est la boule unité de $\mathcal{C}^k(D)$, on a pour l'erreur d'approximation maximale sur $\mathcal{K}$ l'estimation

$$e_n(\mathcal{K})_{L^\infty} \leq Cn^{-k}. \tag{1.20}$$

La décroissance est encore plus rapide pour des fonctions analytiques. Par exemple si $D=[-1,1]$ et que u est dévelopable autour de 0 en série entière de rayon de convergence $R>1$, alors (**exercice**) pour tout $1<r<R$ on a la décroissance géométrique

$$e_n(u)_{L^\infty} \leq Cr^{-n}. \tag{1.21}$$

Si on considère à présent, toujours pour $D=[a,b]$ l'approximation par l'espace V_n des fonctions constantes par morceaux sur la partition uniforme donnée par les intervalles

$$I_k = [a+kh, a+(k+1)h], \quad k=0,\dots,n-1, \quad h := \frac{b-a}{n}, \tag{1.22}$$

alors on montre (**exercice**) que si u est Lipschitzienne alors l'erreur d'approximation pour la norme L^∞ décroit comme n^{-1}. Cependant, on a pas de décroissance plus rapide pour les fonctions plus régulières, même de classe $\mathcal{C}^\infty$, considérer par exemple la fonction $u(x)=x$. Ceci traduit le fait que contrairement aux polynômes, les fonctions constantes par morceaux ont un ordre d'approximation limité (on parle d'une méthode d'ordre 1). On peut bien sûr augmenter l'ordre en considérant des fonctions polynomiales par morceaux, splines, éléments finis.

2. La constante de Lebesgue Λ_n ne dépend pas de u, uniquement du choix de V_n et des points x^i. Si on fixe V_n, le comportement de Λ_n peut fortement varier suivant les points d'échantillonage. Ainsi si V_n est l'espace des polynômes de degrés $n-1$ sur l'intervalle $D=[-1,1]$, le choix le plus évident de points d'interpolation semble être les points équidistants

$$x^i = -1+(i-1)h, \quad h=\frac{2}{n-1}, \tag{1.23}$$

or ce choix conduit à une constante de Lebesgue qui croît plus rapidement que 2^n. Ceci se traduit en particulier par le fait que l'interpolation peut ne pas converger même pour des fonctions analytiques (phénomène de Runge). Un bien meilleur choix est fourni par les points de Chebychev

$$x^i = \cos\Big(\frac{2i+1}{2n+2}\pi\Big), \tag{1.24}$$

pour lesquels on sait que $\Lambda_n = \mathcal{O}(\log n)$. Pour des espaces de polynômes en dimension supérieure à 1 sur des domaines $D\subset\mathbb{R}^d$ généraux, le choix de points

rendant la constante de Lebesgue petite est beaucoup plus difficile et source de problèmes ouverts. Si on considère à présent l'espace V_n des fonctions constantes par morceaux sur une partition $D_1, \ldots, D_n$ d'un domaine $D \subset \mathbb{R}^d$, il est immédiat de voir que le choix d'un point x^i par sous-domaine D_i permet de définir l'opération d'interpolation et que la constante de Lebesgue a la valeur minimale $\Lambda_n = 1$. En ce sens l'interpolation par les fonctions constantes par morceaux est très stable, mais l'ordre de précision de l'approximation est limité comme on l'a expliqué plus haut.

Lorsque les données ponctuelles sont bruitées, c'est à dire de la forme $y^i = u(x^i) + \eta^i$ où $\eta = (\eta^1, \ldots, \eta^n)$ est un vecteur qui représente le bruit de mesure, l'approximation obtenue $\tilde{u} = R_n(y^1, \ldots, y^n)$ vérifie

$$\|\tilde{u} - u\|_{L^\infty} \leq \|u - I_n u\|_{L^\infty} + \|R_n \eta\|_{L^\infty} \leq (1 + \Lambda_n) e_n(u)_V + \Lambda_n \|\eta\|_{\ell^\infty}, \tag{1.25}$$

ce qui montre que la constante de Lebesgue contrôle aussi l'amplification du niveau du bruit dans la reconstruction.

On voit ainsi qu'interpoler des données fortement bruitées (dans le cadre de la régression en particulier) n'est pas très une bonne idée puisque le terme $\Lambda_n \|\eta\|_{\ell^\infty}$ ne tend pas vers 0 quand $n \to \infty$. Dans ce contexte, on cherche à à éviter de sur-ajuster les données en opérant une forme de lissage.

Pour donner un exemple simple, supposons que la fonction u recherchée soit une fonction constante : $u(x) = c \in \mathbb{R}$ pour tout $x \in D$. On observe ainsi $y^i = c + \eta^i$ en chaque point de mesure $x^1, \ldots, x^m$, et on suppose que les η^i sont des réalisation independantes d'une variable η centrée et de variance κ^2, c'est à dire

$$\mathbb{E}(\eta) = 0 \quad \text{et} \quad \mathbb{E}(|\eta|^2) = \kappa^2. \tag{1.26}$$

Si on cherche à interpoler les données en prenant des fonctions constantes par morceaux sur une partition $(D_1, \ldots, D_m)$ telle que x^i appartient à D_i, procédé dont on a vu qu'il était très stable au sens où $\Lambda_m = 1$, alors on voit que la fonction $\tilde{u}$ reconstruite satisfaira

$$\|u - \tilde{u}\|_{L^\infty} = \max_{i=1,\ldots,m} |\eta^i|, \tag{1.27}$$

qui ne tend pas vers 0 quand m augmente. Puisqu'on a supposé ici que u est constante, un meilleur estimateur est donné par la fonction constante obtenue en moyennant les évaluation

$$\tilde{u}(x) = \tilde{c} := \frac{1}{m} \sum_{i=1}^{m} y^i. \tag{1.28}$$

On voit ainsi que $u - \tilde{u} = c - \tilde{c} = \frac{1}{m} \sum_{i=1}^m \eta^i$ qui tend vers 0 en un sens probabiliste : en utilisant le fait que $\mathbb{E}(\eta^i \eta^j) = 0$ si $j \neq i$, on a

$$\mathbb{E}(|c - \tilde{c}|^2) = \frac{\kappa^2}{m}. \tag{1.29}$$

Ceci nous montre l'intérêt de lisser les données fortement bruitées, et le fait qu'on ne peut dans ce cas difficilement espérer de vitesse de convergence plus rapide que m^{-1} au sens de l'erreur quadratique moyenne $\mathbb{E}(|u(x) - \tilde{u}(x)|^2)$.

Pour clore ce chapitre, donnons une première intuition des difficultés qui se posent en grande dimension. On reprend l'exemple de l'approximation par les fonctions constantes par morceaux sur une partition $D_1, \dots, D_n$ de $D \subset \mathbb{R}^d$. Si u est une fonction Lipschitzienne, on l'approche par une fonction $v \in V_n$ qui prend sur chaque D_i une valeur de u en un point de D_i. Ceci entraine (**exercice**) l'estimation

$$\|u - v\|_{L^\infty} \leq Lh, \tag{1.30}$$

où

$$L = \|\nabla u\|_{L^\infty} = \sup_{x \in D} |\nabla u(x)|, \tag{1.31}$$

est la constante de Lipschitz lorsqu'on mesure la distance sur $\mathbb{R}^d$ avec la norme euclidienne $|\cdot| := \|\cdot\|_2$, et

$$h = \max_i \operatorname{diam}(D_i) = \max_i \max_{x, \tilde{x} \in D_i} |x - \tilde{x}|. \tag{1.32}$$

est le diamètre maximal des éléments de la partition. On a donc

$$e_n(u)_{L^\infty} \leq Lh. \tag{1.33}$$

Il est facile de vérifier (**exercice**) que le volume des D_i vérifie

$$|D_i| \leq h^d, \tag{1.34}$$

et comme les D_i forment une partition, on a

$$|D| = \sum_{i=1}^n |D_i| \leq nh^d, \tag{1.35}$$

Ceci nous montre que pour que la partition soit de résolution h, son nombre d'éléments n doit être au moins supérieur à $|D|h^{-d}$ qui explose exponentiellement avec la dimension. Réciproquement puisque D est borné, pour tout $h > 0$, on peut construire une partition par des éléments de diamètre inférieur à h (utiliser par exemple des cubes de longueur $d^{-1/2}h$) et de cardinalité $n \leq Ch^{-d}$ où C est liée au diamètre de D et à la dimension d.

En d'autre terme $h \sim n^{-1/d}$ ce qui nous montre que la borne Lh sur $e_n(u)$ est au mieux de la forme $\mathcal{O}(n^{-1/d})$. Si $\mathcal{K}$ est la boule unité des fonctions lipschitiennes, on obtiendra donc au mieux une estimation de l'erreur maximale du type

$$e_n(\mathcal{K})_{L^\infty} \leq Cn^{-1/d}. \tag{1.36}$$

Cette détérioration de la vitesse de convergence de l'erreur d'approximation avec la dimension d est parfois appelée *malédiction de la dimensionalité* traduction du terme *curse of dimensionality* introduit par Von Neumann. Elle indique ici qu'avec les fonction constantes par morceaux l'approximation de fonctions lipschitziennes à une précision ε donnée exige une partition dont la cardinalité n explose comme ε^{-d} c'est à dire exponentiellement avec la dimension. Nous verrons dans le chapitre suivant que ce phénomène n'est pas lié à l'usage particulier des fonctions constantes par morceaux mais reflette une obstruction plus fondamentale à toutes les méthodes d'approximation. Dans le cas de l'interpolation, signalons que dans de nombreux exemples on observe aussi une déterioration de la constante de Lebesgue Λ_n lorsque la dimension d augmente.

2 Notions de théorie de l'approximation

Nous allons donner dans ce chapitre quelques techniques systématiques permettant d'étudier dans un cadre plus général les quantités $e_n(u)$ et $e_n(\mathcal{K})$ introduites précèdemment. On pourra trouver des approfondissement dans la référence [1] de DeVore citée en introduction.

2.1 Approximation linéaire et non-linéaire

On s'intéresse à l'approximation d'une fonction arbitraire u appartenant à un espace de Banach V de fonctions définies sur un domaine $D \subset \mathbb{R}^d$, par des fonctions "plus simples" choisies dans l'un des éléments d'une famille $(V_n)_{n\geq 0}$, tel que tout élément de V_n peut être décrit par un nombre fini n (ou plus généralement $\mathcal{O}(n)$) de paramètres, c'est à dire par n nombres réels. Pour $n = 0$ on pose simplement

$$V_0 = \{0\}. \tag{2.1}$$

On rappelle la notation

$$e_n(u) = e_n(u)_V := \inf_{v\in V_n} \|v - u\|_V. \tag{2.2}$$

On parle d'approximation *linéaire* lorsque V_n est un espace vectoriel de dimension n (ou $\mathcal{O}(n)$). Quelques exemples élémentaires sont les suivants :

1. Polynômes : pour $D = [a, b] \subset \mathbb{R}$, on considère $V_n = \mathbb{P}_n$ l'espace des polynômes de degré n. qui est de dimension $n + 1$. Un élément $p \in V_n$ est ainsi décrit par $n+1$ paramètres, par exemple ses coefficients polynomiaux, ou ses valeurs en $n+1$ points distincts.
2. Fonction constantes par morceaux : pour $D = [a, b]$, on considère V_n l'espace des fonctions constantes par morceaux sur la partition uniforme $(D_1, \ldots, D_n)$ où $D_j = [a + (j - 1)h, a + jh]$ où $h = (b - a)/n$, qui est de dimension n.
3. Approximation linéaire dans une base de fonctions : si $(\psi_k)_{k\geq 1}$ est une base de l'espace de Banach V, on considère $V_n := \text{vect}\{\psi_1, \ldots, \psi_n\}$.

Bien sûr les exemples 1 et 2 peuvent être étendus au cadre multidimensionel. Pour l'exemple 1, notons que si on considère $V_n = \mathbb{P}_k$ l'ensemble des polynômes de degré total k et de d variables, on a $n = \dim(V_n) = \binom{k+d}{d}$. L'exemple 2 peut se aussi se généraliser en considérant des fonctions polynomiales de degré k prescrit par morceaux, et en imposant des propriétés de régularité aux noeuds $a + jh$ (splines, éléments finis).

Pour l'exemple 3, il convient de préciser la notion de base en dimension infinie : on dit que $(\psi_k)_{k\geq 1}$ est une base de Schauder si et seulement si pour tout $u \in V$ il existe une unique suite $(c_k)_{k\geq 1}$ telle que

$$\lim_{n\to\infty} \Big\| u - \sum_{k=1}^{n} c_k \psi_k \Big\|_V = 0, \tag{2.3}$$

Une notion plus forte est celle de base inconditionelle, qui suppose en plus qu'il existe une constante C telle que pour toutes suites $(c_k)_{k\geq 1}$ et $(d_k)_{k\geq 1}$ à supports finis et tels que $|c_k| \leq |d_k|$ pour tout k, on a

$$\Big\| \sum_k c_k \psi_k \Big\|_V \leq C \Big\| \sum_k d_k \psi_k \Big\|_V. \tag{2.4}$$

Ceci peut-être vu comme une propriété de stabilité : si on effectue des opérations de réduction, seuillage, perturbations des coordonées, on augmente pas trop la norme de la fonction. Cette propriété assure aussi (**exercice difficile**) que la convergence est maintenue en permutant arbitrairement les termes de la série. Si V est un espace de Hilbert, on dit que $(\psi_k)_{k\geq 1}$ est une base de Riesz si et seulement si ses combinaisons linéaires finies sont denses dans V (famille totale) et il existe des constantes $0 < A < B$ telles que toute suite $(c_k)_{k\geq 1}$ à supports fini

$$A \sum_k |c_k|^2 \leq \Big\| \sum_k c_k \psi_k \Big\|_V^2 \leq B \sum_k |c_k|^2. \tag{2.5}$$

Une base Hibertienne est un cas particulier de base de Riesz, qui est un cas particulier de base inconditionelle. La base des séries de Fourier qui est orthonormée dans L^2, n'est pas inconditionelle pour les autres espaces L^p, c'est seulement une base de Schauder. Des exemples de bases inconditionelles dans les espaces L^p et Sobolev sont fournis par les bases d'ondelettes. Dans $\ell^p(\mathbb{N})$ la base canonique des suites de Kroenecker $(\delta_j)_{j\in\mathbb{N}}$ est inconditionelle. En dimension finie, il est facile de vérifier que toute base est inconditionelle pour un choix de norme quelconque, et de Riesz pour une norme Hilbertienne.

On parle d'approximation *non-linéaire* lorsque V_n n'est pas un espace vectoriel - $(v, w) \in V_n$ n'implique pas $v + w \in V_n$ - mais ses éléments peuvent cependant être décrits par $\mathcal{O}(n)$ paramètres. On suppose toujours la propriété d'homogénéité c'est à dire $v \in V_n$ implique $\lambda v \in V_n$ pour $\lambda \in \mathbb{R}$. Les exemples élémentaires suivants peuvent être vus comme des versions non-linéaires des trois exemples linéaires précèdents :

1. Fractions rationnelles : pour $D = [a, b] \subset \mathbb{R}$, on considère $V_n = \{\frac{p}{q} \ : \ p, q \in \mathbb{P}_n\}$ l'ensemble des fractions rationnelles de degré n. qui est caractérisé par $2(n + 1)$ paramètres.
2. Fonctions constantes par morceaux sur des partitions libres : pour $D = [a, b] \subset \mathbb{R}$, on considère V_n l'ensemble de toutes les fonctions v qui sont constantes sur les intervalles $[x_i, x_{i+1}]$ d'une partition $a = x_0 < x_1 < \cdots < x_n = b$ qui peut être dépendre de v. Un élément $v \in V_n$ est ainsi caractérisé par la donnée de $x_1, \ldots, x_{n-1}$ et des valeurs prises sur les $[x_i, x_{i+1}]$, soit $2n - 1$ paramètres.
3. Approximations à n termes dans une base de fonctions : si $(\psi_k)_{k\geq 1}$ est une base de l'espace de Banach V, on considère V_n l'ensemble de toutes les combinaisons linéaires d'au plus n termes, c'est à dire des fonctions de la forme $v = \sum_{k\in E} c_k \psi_k$, avec $\#(E) \leq n$. Un élément $v \in V_n$ est ainsi caractérisé par la donnée des n indices $k_1, \ldots, k_n$ qui constituent E et des n coefficients c_{k_i}.

La non-linéarité de ces exemples est évidente : (i) la somme de deux fractions rationnelles de degré n est en général une fraction rationnelle de degré $2n$ car les dénominateurs

se multiplient, (ii) la somme de deux fonctions constantes par morceaux sur deux partitions différente de cardinal n est en général constante par morceaux sur une partition de cardinal $2n-1$ obtenue en combinant les deux subdivisions, et (iii) la somme de deux combinaisons à n termes est en général une combinaison à $2n$ termes. On voit qu'on a tout de même dans tous ces exemples la propriété $V_n + V_n \subset V_{2n}$. Plus généralement on dit que la famille $(V_n)_{n\geq 0}$ est "faiblement non-linéaire" si il existe une constante $c > 1$ telle que $V_n + V_n \subset V_{cn}$.

Des cas où la famille n'est pas faiblement non-linéaire apparaissent quand on considère les fonctions constantes (ou polynomiales) par morceaux sur des partitions libres en dimension $d > 1$. En prenant par exemple $D = [0,1]^2$ et pour V_n les fonctions constantes par morceaux sur les partitions de D en n rectangles, on vérifie (**exercice**) qu'on a pas mieux que $V_n + V_n \subset V_{n^2}$.

Un autre exemple d'approximation non-linéaire est fourni par les réseaux de neurones, c'est à dire les fonctions $v : \mathbb{R}^d \to \mathbb{R}$ qui peuvent être décrites comme une composition de la forme

$$v = A_m \circ \sigma \circ A_{m-1} \circ \cdots \circ \sigma \circ A_1, \tag{2.6}$$

où pour une suite $d_0, d_1, \ldots, d_m$ avec $d_0 = d$ et $d_m = 1$, les $A_j : \mathbb{R}^{d_{j-1}} \to \mathbb{R}^{d_j}$ sont des applications affines (c'est à dire $(A_j z)_k = \langle a_{j,k}, z\rangle + b_{j,k}$ avec $a_{j,k} \in \mathbb{R}^{d_{j-1}}$ et $b_{j,k} \in \mathbb{R}$) et σ est une fonction d'activation qu'on applique composante par composante sur un vecteur $z = (z_1, \ldots, z_l)^T$ en posant $\sigma(z) := (\sigma(z_1), \ldots, \sigma(z_l))^T$. Un choix de σ souvent utilisé est la fonction RELU $\sigma(t) = \max(0, t)$. On peut ainsi désigner par V_n l'ensemble des fonctions de ce type lorsque le nombre total de paramètres entrant en jeu (la suite des dimension d_j et les coefficients des applications A_j) est inférieur à n.

Tous les exemples qu'on a présenté font apparaitre des familles $(V_n)_{n\geq 0}$, permettant d'améliorer la précision de l'approximation lorsque n augmente. En particulier $(e_n(u))_{n\geq 0}$ est une suite décroissante si l'on suppose la propriété d'emboitement $V_n \subset V_{n+1}$. Comme on l'a déjà remarqué l'infimum dans la définition de l'erreur de meilleure approximation $e_n(u)_V$ est toujours atteint dans le cas linéaire, c'est à dire qu'il existe un $u_n \in V_n$ tel que $\|u - u_n\|_V \leq \|u - v\|_V$ pour tout $v \in V_n$. Ceci n'est plus toujours vrai dans le cas non-linéaire.

Un problème central en théorie de l'approximation est de *caractériser* par une propriété simple (typiquement de régularité) les fonctions u telles que $e_n(u)_V$ décroit à une certaine vitesse, lorsqu'on s'est fixé la famille $(V_n)_{n\geq 0}$ et la norme $\|\cdot\|_V$ dans laquelle on mesure l'erreur. On peut ainsi définir pour tout $s > 0$ l'espace

$$\mathcal{A}^s = \mathcal{A}^s((V_n)_{n\geq 0}, V) := \{u \in V \ : \ \sup_{n\geq 0} n^s e_n(u)_V < \infty\}, \tag{2.7}$$

c'est à dire l'ensemble des fonctions de V telles que $e_n(u)_V$ décroit à la vitesse $\mathcal{O}(n^{-s})$. On vérifie qu'il s'agit bien d'un espace vectoriel (**exercice**) dans le cas où $(V_n)_{n\geq 0}$ est une famille d'espaces linéaires, ou une famille faiblement non-linéaire. La quantité

$$\|u\|_{\mathcal{A}^s} := \|u\|_V + \sup_{n\geq 0} n^s e_n(u)_V, \tag{2.8}$$

constitue alors une quasi-norme pour $\mathcal{A}^s$: elle a toutes les propriétés d'une norme mis à part l'inégalité triangulaire qui doit être remplacée par $\|u+v\| \leq C(\|u\| + \|v\|)$ où C est

une constante fixée. En revanche pour des familles fortement non-linéaires, il est possible que $v, w \in \mathcal{A}^s$ n'implique pas $v + w \in \mathcal{A}^s$ (**exercice** : considérer l'exemple donné plus haut des fonctions constantes par morceaux sur des rectangles et les fonctions $v(x, y) = x^2$ et $w(x, y) = y^2$). Une question centrale est donc de caractériser l'appartenance de u à $\mathcal{A}^s$ par une propriété analytique simple (régularité, localisation, intégrabilité) qui puisse se vérifier sans avoir examiner le comportement de $e_n(u)$ pour tout n. Nous donnerons plus loin quelques exemples de réponses à ce problème qui permettront d'illustrer les différences entre l'approximation linéaire et non-linéaire.

Un autre problème central est la construction pratique d'approximations d'une fonction u dans V_n. Le minimiseur exact de $\|u - v\|_V$ sur tous les $v \in V_n$ est souvent difficile à caractériser et calculer. Un cas simple est celui où V_n est un sous-espace vectoriel d'un espace de Hilbert V. Dans ce cas le minimiseur est donné par la projection orthogonale de u sur V_n. Si l'on dispose d'une base orthonormée $\{L_1, \ldots, L_n\}$ de V_n, celle-ci se calcule suivant

$$P_n u = \sum_{j=1}^{n} \langle u, L_j \rangle L_j. \tag{2.9}$$

Si la base n'est pas orthonormée, le calcul de $P_n u = \sum_j a_j L_j$ reste élémentaire : en utilisant les équations $\langle u - P_n u, L_j \rangle = 0$ pour tout j, on voit que le vecteur $\mathbf{a} = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ est solution du système

$$\mathbf{G}\mathbf{a} = \mathbf{b}, \tag{2.10}$$

où $\mathbf{G} = (\langle L_j, L_i \rangle)_{i,j=1,\ldots,n}$ est la matrice de Gramm et $\mathbf{b} \in \mathbb{R}^n$ est le vecteur de coordonnées $\langle u, L_j \rangle$.

Dans le cas où V n'est pas un espace de Hilbert, la caractérisation du minimiseur n'est plus claire : même si V_n est un espace linéaire, l'application qui à u associe sa meilleure approximation dans V_n n'est en général pas linéaire et n'a pas d'expression simple à quelques rares exceptions près. Un objectif plus raisonnable est la mise au point de méthodes d'approximation $u \mapsto u_n \in V_n$ calculables en pratique et quasi-optimales au sens où il existe une constante $C \geq 1$ telle que pour tout $u \in V$ et tout $n \geq 0$, on a

$$\|u - u_n\|_V \leq C e_n(u)_V. \tag{2.11}$$

On remarque que si P_n est un opérateur de projection sur V_n borné au sens de la norme V, on a pour tout $v \in V_n$

$$\|u - P_n u\|_V \leq \|u - v\|_V + \|P_n(u - v)\|_V \leq (1 + \|P_n\|_{V \to V})\|u - v\|. \tag{2.12}$$

Ainsi, si l'on dispose d'opérateurs de projection sur V_n uniformément stables au sens où $\|P_n\|_{V \to V} \leq D$ indépendamment de n, alors les approximations $u_n = P_n u$ satisfont la quasi-optimalité avec $C = 1 + D$.

L'existence de projecteurs uniformément bornés dans les espaces de Banach n'est en général pas garantie : un célèbre théorème du à Kadec et Snobar affirme que pour un espace V_n quelconque de dimension n dans un espace de Banach général, on peut au mieux obtenir un projecteur de norme $\|P_n\|_{V \to V} \leq \sqrt{n}$. Cette borne s'améliore pour des espaces de Banach plus spécifiques (elle est en n^{r_p} avec $r_p = \max\{0, 1/2 - 1/p\}$ dans le cas des espaces L^p) et elle peut être bien meilleure pour des espaces V_n particuliers.

Par exemple, si on considère les espaces V_n de fonction constantes par morceaux sur une partition $(D_1, \dots, D_n)$ on montre facilement que la projection L^2-orthogonale est de norme 1 dans L^p (**exercice**). Un autre exemple est la projection L^2 orthogonale sur les premiers éléments de la base des séries de Fourier, c'est à dire la somme partielle de Fourier $S_n u$ pour laquelle on montre (**exercice**) que $\|S_n\|_{L^\infty \to L^\infty} \sim \log(n)$.

La construction d'approximations quasi-optimales devient plus difficile dans le cas de familles $(V_n)_{n\geq 0}$ non-linéaires. Les procédés d'approximation non-linéaires ont par nature un caractère *adaptatif*. Par exemple, dans l'exemple de fonctions constantes ou polynomiales par morceaux sur des partitions libres, il s'agit de comprendre pour une fonction u donnée quelle est la partition la mieux adaptée à son approximation. Nous donnerons un exemple simple dans la section suivante. Un cas important où l'approximation optimale peut être caractérisée et calculée est celui des combinaison à n termes dans une base orthonormée $(\psi_k)_{k\geq 1}$, lorsque V est un espace de Hilbert. En effet si $u = \sum_{k\geq 1} c_k \psi_k$ est un élément de V et $v = \sum_{k\in E} d_k \psi_k$ est une combinaison à $n = \#(E)$ termes, on a, par l'égalité de Parseval,

$$\|u - v\|_V^2 = \sum_{k\in E} |c_k - d_k|^2 + \sum_{k\notin E} |c_k|^2. \tag{2.13}$$

Cette quantité est minimisée en choisissant $d_k = c_k$ pour $k \in E$, et en prenant l'ensemble $E = E_n(u)$ de cardinal n qui minimise $\sum_{k\notin E} |c_k|^2$, c'est à dire l'ensemble des indices correspondant aux n plus grandes coordonnées $|c_k|$ de u (noter que cet ensemble peut ne pas être unique en cas d'égalité en valeur absolue de certaines coordonnées). L'approximation obtenue $u_n = \sum_{k\in E_n(u)} c_k \psi_k \in V_n$ vérifie ainsi

$$\|u - u_n\|_V = e_n(u)_V = (\sum_{k\notin E} |c_k|^2)^{1/2}. \tag{2.14}$$

Notons que cette approximation correspond essentiellement à opérer un seuillage des coordonnées de u en gardant uniquement celles qui sont supérieures à une valeur donnée et en mettant les autres à 0. On parle parfois d'approximation *parcimonieuse* (sparse approximation), car cela revient à chercher la meilleure approximation dans ℓ^2 du vecteur des coordonnées $(c_k)_{k\geq 1}$ par un vecteur dont le nombre de coordonnées non-nulle est au plus n. Le caractère adaptatif du procédé se traduit par le fait que l'ensemble $E = E_n(u)$ dépend de la fonction u que l'on approche, à l'inverse du cas linéaire où l'on prend $E = \{1, \dots, n\}$ pour toute fonction u. On peut montrer (**exercice**) que ce même processus conduit à une approximation quasi-optimale dans le cas d'une base de Riesz.

L'approximation parcimonieuse joue un rôle important dans de multiples applications. Citons en particulier la compression d'image avec le standard JPEG 2000 qui utilise des bases de Riesz d'ondelettes dites biorthogonales pour représenter les images numériques et allouer l'essentiel des bits sur les plus grands coefficients. Le développement récent de méthodes du *Compressed Sensing* pour résoudre des problèmes linéaires sous-déterminés est aussi fondé sur l'idée de parcimonie : rechercher la solution de plus petit support d'un système $Ax = b$ de taille $n \times N$ avec $N >> n$ qui admet donc un nombre infini de solutions.

2.2 Vitesse d'approximation : régularité et sommabilité

Nous allons étudier les vitesses d'approximation sur deux exemples élémentaires mais importants qui permettent de comparer les méthodes linéaires et non-linéaires. Notre étude fera intervenir des espaces de fonctions régulières et des espaces de suites. On suppose ici les étudiants familiers avec les espaces classiques de fonctions (qui sont tous des espaces de Banach pour la norme qui leur correspond) : les espaces $\mathcal{C}^k(D)$ des fonctions k fois continuement dérivables, l'espace $\mathrm{Lip}(D)$ des fonctions lipschitziennes, les espaces $\mathcal{C}^\alpha(D)$ des fonctions α-hölderiennes pour $0 < \alpha < 1$, et les espaces de Sobolev $W^{s,p}(D)$ qui contient les fonctions dont les dérivées partielles jusqu'à l'ordre s (au sens des distributions) sont dans $L^p(D)$ et qui sont aussi notés $H^s(D)$ lorsque $p = 2$.

On étudie tout d'abord l'approximation de fonctions continues par les fonctions constantes par morceaux en dimension 1, en se placant pour simplifier les notations sur l'intervalle unité $D = [0, 1]$. On se place donc dans $V := \mathcal{C}(D)$ que l'on munit de la norme L^∞. On fait ici une légère entorse au cadre théorique qu'on s'est fixé : les espaces V_n de fonctions constantes par morceaux ne sont pas contenus dans $\mathcal{C}(D)$, ils sont cependant dans L^∞ et on peut ainsi mesurer l'erreur $\|u - v\|_{L^\infty}$ pour tout $u \in V$ et $v \in V_n$, et donner un sens à $e_n(u)_{L^\infty}$.

Si $(D_1, \ldots, D_n)$ est une partition quelconque de D, il est naturel d'approcher une fonction $u \in V$ par une fonction u_n prenant sur chaque intervalle D_i la valeur constante $u(x_i)$ où x_i est un point choisi arbitrairement dans D_i. On a alors pour tout $i = 1, \ldots, n$ et $x \in D_i$ l'estimation

$$|u(x) - u_n(x)| = |u(x) - u(x_i)| \leq \max_{x,y \in D_i} |u(x) - u(y)| = \max_{x \in D_i} u(x) - \min_{x \in D_i} u(x) \qquad (2.15)$$

Si on veut affiner, on peut prendre sur chaque D_i la constante c qui minimise l'erreur locale $\|u - c\|_{L^\infty(D_i)}$ c'est à dire la valeur médiane $c = \frac{1}{2}(\max_{x \in D_i} u(x) + \min_{x \in D_i} u(x))$ qui par continuité est une valeur prise par u en un point $x_i \in D_i$ et dans ce cas on gagne une constante multiplicative $1/2$ dans l'estimation çi-dessus.

Considérons tout d'abord le cas de partition $(D_1, \ldots, D_n)$ uniformes, c'est à dire

$$D_i = [(i-1)/n, i/n]. \qquad (2.16)$$

On note V_n l'espace vectoriel des fonctions constantes par morceaux sur cette partition, et on a ainsi une famille $(V_n)_{n \geq 1}$ d'approximation linéaire. Afin d'étudier la vitesse de décroissance de $e_n(u)$, supposons que u soit lipschitzienne, ce qui est équivalent à dire que $u' \in L^\infty$ où $L = \|u'\|_{L^\infty}$ est la constante de Lipschitz de u. On a alors l'estimation

$$\max_{x,y \in D_i} |u(x) - u(y)| \leq L|D_i| = \|u'\|_{L^\infty} n^{-1}, \qquad (2.17)$$

ce qui nous montre que la régularité lipchitzienne de u entraîne la décroissance

$$e_n(u)_{L^\infty} \leq \|u - u_n\|_{L^\infty} \leq C n^{-1}, \quad n \geq 1, \qquad (2.18)$$

avec $C = \|u'\|_{L^\infty}$. On peut ici prouver une propriété réciproque : soit $u \in \mathcal{C}(D)$ telle que l'estimation de décroissance çi-dessus soit satisfaite pour une certaine constante C. Pour montrer que cette fonction est lipschitizienne, on prend deux points quelconques $x, y \in D$.

Pour tout $n \geq 1$, on sait qu'il existe $u_n \in V_n$ telle que $\|u - u_n\|_{L^\infty} \leq Cn^{-1}$ et on peut ainsi écrire

$$|u(x) - u(y)| \leq |u_n(x) - u_n(y)| + 2Cn^{-1}. \tag{2.19}$$

On choisit $n \geq 1$ tel que $\frac{1}{2n} \leq |x - y| \leq \frac{1}{n}$, et on note $(D_1, \ldots, D_n)$ la partition correspondante. Deux possibilités : (i) x et y appartiennent au même intervalle D_i auquel cas $|u_n(x) - u_n(y)| = 0$ ou (ii) x et y appartiennent à deux intervalles adjacents D_i et D_{i+1}. En notant $z = ih$ le point de raccord entre ces intervalles, on peut alors écrire, grace à la continuité de u,

$$|u_n(x) - u_n(y)| \leq |u_n(x) - u(z)| + |u(z) - u_n(y)| \leq 2\|u - u_n\|_{L^\infty} \leq 2Cn^{-1}. \tag{2.20}$$

On a ainsi obtenu dans tous les cas

$$|u(x) - u(y)| \leq 4Cn^{-1} \leq 8C|x - y|, \tag{2.21}$$

ce qui prouve que u est lipschitzienne avec $\|u'\|_{L^\infty} \leq 8C$. En résumé on a identifié la classe de fonctions approchable à la vitesse n^{-1} comme celle des fonctions lipschitziennes, ce qui se traduit par le résultat suivant.

Théorème 2.1 *Pour $V = \mathcal{C}(D)$ avec $D = [0, 1]$ et $(V_n)_{n\geq 0}$ la famille des fonctions constantes par morceaux sur des partitions uniformes, on a*

$$\mathcal{A}^1 = \mathrm{Lip}(D). \tag{2.22}$$

Quelques généralisations sont immédiates : si u est une fonction s-hölderienne pour un $0 < s < 1$, ce qui signifie que $|u(x) - u(y)| \leq C|x - y|^s$, les mêmes raisonnements montrent que $e_n(u)_{L^\infty} \leq Cn^{-s}$ et qu'on a une implication réciproque. En d'autre termes

$$\mathcal{A}^s = \mathcal{C}^s(D), \quad 0 < s < 1. \tag{2.23}$$

En dimension supérieure, si $D \subset \mathbb{R}^d$ est un domaine borné, on peut introduire des suites de partitions uniformes, par exemple en considérant l'intersection de D par une partition de $\mathbb{R}^d$ en cubes uniformes. Comme on l'a déjà remarqué, on obtiendra pour u lipschitzienne sur D la vitesse $e_n(u) \leq Cn^{-1/d}$ et un raisonnement similaire permet d'obtenir une réciproque. De même pour les fonctions s-hölderiennes avec la vitesse $n^{-s/d}$. En d'autres termes $\mathcal{A}^{1/d} = \mathrm{Lip}(D)$ et $\mathcal{A}^{s/d} = \mathcal{C}^s(D)$ pour $0 < s < 1$. Enfin comme on la déjà observé, la vitesse ne sera pas améliorée pour des classe de fonctions plus régulières, ce qui traduit le caractère limité en ordre de l'approximation par les fonctions constantes par morceaux.

Considérons à présent le cas de la famille non-linéaire $(V_n)_{n\geq 0}$ des fonctions constantes par morceaux sur des partitions libres : pour construire l'approximation $u_n \in V_n$ on peut choisir la partition $(D_1, \ldots, D_n)$ en fonction de la fonction u que l'on cherche à approcher. Supposons cette fois que la fonction u appartienne à l'espace de Sobolev $W^{1,1}(D)$ c'est à dire que $u' \in L^1(D)$. Pour tout i, on peut cette fois utiliser la majoration

$$\max_{x,y\in D_i} |u(x) - u(y)| \leq \int_{D_i} |u'(t)|dt. \tag{2.24}$$

Il est alors naturel de choisir une partition qui *équidistribue* la norme L^1 de u' c'est à dire telle que

$$\int_{D_i} |u'(t)|dt = n^{-1} \int_D |u'(t)|dt, \tag{2.25}$$

et on obtient ainsi

$$e_n(u)_{L^\infty} \leq \|u - u_n\|_{L^\infty} \leq Cn^{-1}, \tag{2.26}$$

avec $C = \|u'\|_{L^1}$. Il est là aussi possible (**exercice difficile**) d'établir une réciproque : si $e_n(u)_{L^\infty}$ vérifie la décroissance çi-dessus, alors $u \in W^{1,1}(D)$. Ceci se résume par le résultat suivant.

Théorème 2.2 *Pour $V = \mathcal{C}(D)$ avec $D = [0,1]$ et $(V_n)_{n\geq 0}$ la famille des fonctions constantes par morceaux sur des partitions libres, on a*

$$\mathcal{A}^1 = W^{1,1}(D). \tag{2.27}$$

Ainsi la vitesse Cn^{-1} est atteinte sous une condition significativement plus faible que pour l'approximation linéaire : $u' \in L^1$ plutôt que $u' \in L^\infty$. Autrement dit on a agrandi l'ensemble $\mathcal{A}^1$ des fonctions approchables à la vitesse n^{-1}, ce qui signifie que certaines fonctions seront approchables à cette vitesse avec des partitions adaptatives mais pas avec des partitions uniformes. Prenons par exemple pour $0 < s < 1$ la fonction

$$u(x) = x^s. \tag{2.28}$$

On voit que $u'(x) = sx^{s-1}$ est intégrable sur $[0,1]$ ce qui montre que $e_n(u)_{L^\infty} \leq Cn^{-1}$ avec des partitions adaptatives. Mais pour des partitions uniformes, on aura au mieux la borne $e_n(u)_{L^\infty} \leq Cn^{-s}$ - inspecter la variation de u sur le premier intervalle $D_1 = [0, \frac{1}{n}]$ - ce qui traduit le fait que la régularité hölderienne de cette fonction n'est pas meilleure que $\mathcal{C}^s$.

Remarque 2.1 *L'exemple qu'on vient de décrire s'insère dans une théorie plus générale d'approximation par des fonctions polynomiales de degré k par morceaux sur un domaine $D \subset \mathbb{R}^d$. On peut résumer les résultats de cette théorie de la façon suivante lorsque l'on mesure l'erreur dans $V = L^p(D)$. Pour $0 < s < k+1$, la vitesse d'approximation $e_n(u)_{L^p} = \mathcal{O}(n^{-s/d})$ est atteinte pour les fonctions u de $W^{s,p}(D)$ si on utilise des partitions uniformes. Si l'on utilise des partitions adaptatives, la même vitesse est atteinte pour les fonctions de $W^{s,q}(D)$ avec*

$$\frac{1}{q} = \frac{1}{p} + \frac{s}{d}. \tag{2.29}$$

Comme $q < p$, l'espace $W^{s,q}(D)$ est strictement plus grand que $W^{s,p}(D)$. L'exemple simple que nous avons traité correspond à $p = \infty$ et $s = d = q = 1$. On pourra noter que la relation çi-dessus correspond à l'injection de Sobolev critique de $W^{s,q}$ dans L^p qui n'est pas compacte (contrairement à celle de $W^{s,p}$ dans L^p qui l'est par le théorème de Rellich). Ceci implique que n'importe quelle méthode linéaire est dans l'incapacité d'approcher les fonctions de $W^{s,q}$ dans L^p, car cela exigerait la compacité de l'injection. Noter aussi que lorsque s augmente, l'exposant q devient plus petit que 1 et l'espace $W^{s,q}(D)$ n'est alors pas bien défini, il faut en fait le remplacer par une variante appelée espace de Besov $B^s_{q,q}(D)$ pour que la théorie garde un sens.

Comme on l'a déjà mentionné, la recherche d'une procédure d'approximation $u \mapsto u_n \in V_n$ ayant une propriété de quasi-optimalité $\|u - u_n\|_V \leq Ce_n(u)_V$ au sens n'est pas simple dans le cas de l'approximation non-lineaire par des partitions adaptatives. On peut cependant donner une idée intuitive et constructive de la partition optimale qui équilibre la norme L^1 de u' dans le cas particulier où u est une fonction monotone, par exemple strictement croissante. Celle-ci revient prendre l'intervalle des valeurs $[u(0), u(1)]$ prises par u et à le découper de manière uniforme sur l'axe des ordonnées en n intervalles

$$I_k = [u(0) + (k-1)h, u(0) + kh], \quad h := (u(1) - u(0))/n, \quad k = 1, \ldots, n. \tag{2.30}$$

L'intervalle $D_k = [x_{k-1}, x_k]$ est défini sur l'axe des abscisses comme l'image réciproque $u^{-1}(I_k)$ c'est à dire tel que $u(x_k) = u(0)+kh$. Il est dans ce cas facile de voir (**exercice**) que la fonction u_n qui vaut la valeur médiane $u(0)+(k-1/2)h$ de I_k sur D_k est l'approximation optimale de u dans V_n.

Remarque 2.2 *On voit ainsi que la partition optimale est aussi celle qui équilibre les erreur d'approximation locale $\|u - u_n\|_{L^\infty(D_k)}$ ici égale à $h/2 = Cn^{-1}$ avec $C = (u(1) - u(0))/2$. Le principe d'équilibrage de l'erreur locale en norme L^∞ (ou plus généralement en norme L^p ou de Sobolev) est au coeur d'algorithmes généraux de raffinement adaptatif : on part d'une partition grossière et on se donne pour objectif une erreur locale $\varepsilon > 0$ fixée. On raffine itérativement les éléments de la partition tant que l'erreur d'approximation locale sur ceux-ci excède $\varepsilon > 0$. Dans le cas de $D = [0, 1]$, on peut ainsi prendre cet intervalle tout entier comme partition initiale et générer une suite de partitions adaptatives en coupant en deux par le milieu tout intervalle I de la partition courante tant que $\min_{c\in\mathbb{R}} \|u-c\|_{L^\infty(I)} \geq \varepsilon$. On peut par exemple choisir de découper à chaque étape l'intervalle où l'erreur locale est la plus grande. L'algorithme se termine avec une partition $(D_1, \ldots, D_n)$ avec $n = n(\varepsilon)$, et une fonction constante par morceaux $u_n \in V_n$ qui approche avec la précision $\|u - u_n\|_{L^\infty} \leq \varepsilon$. Dans un tel procédé, les partitions ne sont plus complètement arbitraires puisque les intervalles D_i ont nécessairement une forme dyadique $[2^{-j}k, 2^{-j}(k+1)]$ avec $j \in \mathbb{N}$ et $k \in \{0, \ldots, 2^j - 1\}$. A cause de cette restriction, on ne peut plus montrer que $\|u - u_n\|_{L^\infty} \leq Cn^{-1}$ lorsque $u \in W^{1,1}(D)$, mais il est possible (***exercice difficile***) de montrer que cette vitesse reste atteinte lorsque $u \in W^{1,p}(D)$ dès que $p > 1$.*

Nous allons à présent illustrer les différences entre les approches linéaires et non-linéaires pour l'approximation dans une base orthonormée $(\psi_k)_{k\geq 1}$ d'un espace de Hilbert V. Toute fonction $u \in V$ se décompose suivant

$$u = \sum_{k\geq 1} c_k \psi_k, \quad c_k = \langle u, \psi_k \rangle. \tag{2.31}$$

Pour l'approximation linéaire, on considère les espaces $V_n = \text{vect}\{\psi_1, \ldots, \psi_n\}$. La meilleure approximation est fournie par la projection orthogonale

$$u_n = P_{V_n} u = \sum_{k=1}^{n} c_k \psi_k. \tag{2.32}$$

L'erreur de meilleure approximation est donnée par

$$e_n(u)_V = \|u - u_n\|_V = \Big(\sum_{k>n} |c_k|^2\Big)^{1/2}. \tag{2.33}$$

La décroissance de $e_n(u)_V$ peut ainsi se relier au la vitesse de convergence de la série $\sum |c_k|^2$. Une manière de quantifier celle-ci est de supposer que cette série converge quand on introduit des poids multiplicatifs qui augmentent avec n. Ainsi, pour $s > 0$ on peut introduire les espaces

$$V^s = \Big\{u \in V \ : \ \sum_{k\geq 1} k^{2s}|c_k|^2 < \infty\Big\}, \tag{2.34}$$

que l'on peut munir de la norme Hilbertienne $\|u\|_{V^s}^2 = \sum_{k\geq 1} k^{2s}|c_k|^2$ qui est une norme ℓ^2 à poids sur la suite des coordonnées de u. Noter que $V^s \subset V^t$ pour $t \leq s$. Pour tout $u \in V^s$, on peut écrire

$$e_n(u)_V^2 = \sum_{k>n} |c_k|^2 \leq (n+1)^{-2s} \sum_{k>n} k^{2s}|c_k|^2 \leq (n+1)^{-2s}\|u\|_{V^s}^2, \tag{2.35}$$

ce qui nous montre que

$$e_n(u)_V \leq Cn^{-s} \tag{2.36}$$

avec $C = \|u\|_{V^s}$, c'est à dire $u \in \mathcal{A}^s$. Contrairement au cas des fonctions constantes par morceaux on ne peut pas prouver une réciproque exacte, mais celle-ci est presque vérifiée au sens suivant : si $u \in \mathcal{A}^s$ alors pour tout $0 < t < s$ on a $u \in V^t$. En effet on peut écrire

$$\sum_{k\geq 1} k^{2t}|c_k|^2 = |c_1|^2 + \sum_{j\geq 0} \sum_{2^j<k\leq 2^{j+1}} k^{2t}|c_k|^2 \leq |c_1|^2 + \sum_{j\geq 0} 2^{2(j+1)t} \sum_{k>2^j} |c_k|^2, \tag{2.37}$$

et par conséquent puisque $\sum_{k>n} |c_k|^2 \leq Cn^{-2s}$,

$$\sum_{k\geq 1} k^{2t}|c_k|^2 \leq |c_1|^2 + C\sum_{j\geq 0} 2^{2(j+1)t}2^{-2js} = |c_1|^2 + 2^{2t}C\sum_{j\geq 0} 2^{-2j(s-t)} < \infty, \tag{2.38}$$

ce qui nous montre que $u \in V^t$. Pour vérifier qu'il n'y pas de réciproque exacte, on pourra construire une fonction $u \in \mathcal{A}^s$ qui n'appartient pas à V^s par un choix particulier d'une suite de coefficients $(c_n)_{n\geq 0}$ (**exercice**). En résumé nous avons montré le résultat suivant.

Théorème 2.3 *Pour l'approximation linéaire dans une base orthonormée d'un espace de Hilbert V, on a pour tout $s > 0$ et $\varepsilon > 0$,*

$$V^s \subset \mathcal{A}^s \quad \text{et} \quad \mathcal{A}^s \subset V^{s-\varepsilon}. \tag{2.39}$$

Examinons la nature des espaces V^s dans le cas particulier où V est l'espace $L^2(D)$ avec $D = [0, 1]$ muni de la base des séries de Fourier

$$\varphi_k(x) = e^{i2\pi kx}, \quad k \in \mathbb{Z}. \tag{2.40}$$

Toute fonction $u \in L^2(D)$ se décompose dans cette base suivant

$$u = \sum_{k\in\mathbb{Z}} d_k \varphi_k, \quad d_k = d_k(u) = \langle u, \varphi_k \rangle = \int_0^1 u(x) e^{-i2\pi kx} dx. \tag{2.41}$$

La base orthonormée $(\varphi_k)_{k\in\mathbb{Z}}$ est ici indexée par $\mathbb{Z}$, et il est ici plus naturel de prendre pour V_n l'espace des polynômes trigonométriques de degré n, c'est à dire

$$V_n := \mathrm{vect}\{\varphi_k \ : \ |k| \leq n\} \tag{2.42}$$

qui est de dimension $2n+1$. La projection de u sur V_n est sa somme de Fourier partielle

$$P_n u = S_n u = \sum_{|k|\leq n} \langle u, \varphi_k \rangle \varphi_k, \tag{2.43}$$

et l'erreur d'approximation est donc donnée par

$$e_n(u)^2_{L^2} = \sum_{|k|>n} |d_k|^2. \tag{2.44}$$

Pour décrire sa décroissance, il faut aussi redéfinir les espaces V^s comme les fonctions $u \in L^2(D)$ telles que

$$\sum_{k\in\mathbb{Z}} |k|^{2s} |d_k|^2 < \infty. \tag{2.45}$$

Les arguments précèdents s'adaptent immédiatement, et on obtient les mêmes inclusions $V^s \subset \mathcal{A}^s \subset V^{s-\varepsilon}$.

Pour $s > 0$ entier, l'espace V^s s'identifie à l'espace de Sobolev périodique $H^s_{per}(D) = W^{s,2}_{per}(D)$. Sans rentrer dans la définition détaillée de cet espace, donnons en l'intuition : si u et ses dérivées jusqu'à l'ordre $s-1$ sont périodiques, alors pour $k \neq 0$ le coefficient de Fourier $d_k(u)$ peut se relier à celui de ses dérivées en itérant l'intégration par partie

$$d_k(u) = \int_0^1 u(x) e^{-i2\pi kx} dx = \frac{1}{i2\pi k} \int_0^1 u'(x) e^{-i2\pi kx} dx = \frac{1}{i2\pi k} d_k(u'), \tag{2.46}$$

puisque les termes de bord disparaissent par périodicité, ce qui mène aux identités

$$d_k(u^{(l)}) = (i2\pi k)^l d_k(u), \quad l = 1, \ldots, s. \tag{2.47}$$

Pour de telle fonctions périodiques, la norme de l'espace $H^s_{per}(D)$ est donnée par

$$\|u\|^2_{H^s_{per}} = \sum_{l=0}^{s} \|u^{(l)}\|^2_{L^2(D)}, \tag{2.48}$$

et l'identité de Parseval donne ainsi la forme d'une norme ℓ^2 à poids

$$\|u\|^2_{H^s_{per}} = \sum_{k\in\mathbb{Z}} \omega_k |d_k(u)|^2, \quad \omega_k = 1 + |2\pi k|^2 + \cdots + |2\pi k|^{2s}. \tag{2.49}$$

Noter l'importance de la périodicité de u et de ses dérivées pour établir que cette somme converge : la fonction $u(x) = x$, bien qu'elle soit $\mathcal{C}^\infty$ sur $[0,1]$ est discontinue quand on la périodise et il est facile de vérifier qu'elle n'appartient pas à $H^s_{per}(D)$ pour $s \geq 1$.

Le poids dominant dans l'expression çi-dessus étant de la forme $C|k|^s$, il est immédiat que $H^s_{per}(D)$ se décrit plus simplement comme l'ensemble des fonctions de $L^2(D)$ telles que $\sum_{k\in\mathbb{Z}} |k|^{2s}|d_k|^2 < \infty$. Ceci permet en particulier d'étendre la définition de cet espace de Sobolev au cas d'une régularité fractionnaire $s \notin \mathbb{N}$. On a ainsi le résultat suivant.

Théorème 2.4 *Pour l'approximation linéaire dans la base des séries de Fourier de $L^2(D)$, on a pour tout $s > 0$ et $\varepsilon > 0$,*

$$H^s_{per}(D) \subset \mathcal{A}^s \quad \text{et} \quad \mathcal{A}^s \subset H^{s-\varepsilon}_{per}(D). \tag{2.50}$$

Remarque 2.3 *On peut réorganiser la base de Fourier en une base $(\psi_n)_{n\geq 1}$, en posant par exemple*

$$\psi_1 = \varphi_0, \quad \psi_{2k} = \varphi_k, \quad \psi_{2k+1} = \varphi_{-k}, \quad k \geq 1. \tag{2.51}$$

On vérifie alors aisément que $u = \sum_{n\geq 1} c_k\varphi_k$ appartient à $H^s_{per}(D)$ si et seulement si $\sum_{n\geq 1} n^{2s}|c_n|^2 < \infty$, et on retrouve ainsi la première définition de l'espace V^s.

Remarque 2.4 *En plusieurs dimension, avec $D = [0,1]^d$, les bases de Fourier sont de la forme*

$$\varphi_k(x) = \exp(i2\pi\langle k, x\rangle) = \exp(i2\pi(k_1x_1 + \cdots + k_dx_d)), \tag{2.52}$$

pour $x = (x_1, \cdots, x_d) \in D$ avec $k = (k_1, \ldots, k_d) \in \mathbb{Z}^d$. On considère alors les espaces

$$V_n = \text{vect}\{\varphi_k \ : \ |k|_\infty = \max|k_i| \leq m\}, \tag{2.53}$$

*qui sont de dimension $n = (2m+1)^d$ et on peut généraliser l'analyse qui a été faite en dimension 1. On obtient par les mêmes arguments (***exercice***) que la décroissance de $e_n(u)_{L^2}$ à la vitesse m^{-s} c'est à dire $n^{-s/d}$ est assurée pour $u \in H^s_{per}(D)$ et plus précisément*

$$H^s_{per}(D) \subset \mathcal{A}^{s/d} \quad \text{et} \quad \mathcal{A}^{s/d} \subset H^{s-\varepsilon}_{per}(D). \tag{2.54}$$

Ce résultat est du même type que le résultat général indiqué dans la remarque 2.1 pour l'approximation par des fonctions polynomiales par morceaux sur des partitions uniformes.

On revient à une base orthonormée $(\psi_k)_{k\geq 1}$ d'un espace de Hilbert général V et on étudie à présent l'approximation non-linéaire à n termes. Nous avons vu que si $u = \sum_{k\geq 1} c_k\psi_k$, l'erreur de meilleure approximation est donnée par

$$e_n(u)^2_V = \sum_{k\notin E_n} |c_k|^2, \tag{2.55}$$

où $E_n = E_n(u)$ est l'ensemble des indices correspondant aux n plus grands $|c_k|$. Afin d'étudier sa décroissance, il est utile d'introduire le *réarrangement décroissant*, c'est à dire la suite $(c^*_k)_{k\geq 1}$ obtenue par permutation des $(|c_k|)_{k\geq 1}$ (il existe une bijection $\pi : \mathbb{N} \to \mathbb{N}$ telle que $c^*_k = |c_{\pi(k)}|$) et telle que

$$c^*_1 \geq c^*_2 \geq \cdots \geq c^*_k \geq \cdots \geq 0. \tag{2.56}$$

Les n plus grandes valeurs de $|c_k|$ sont donc les valeurs $c_1^*, \ldots, c_n^*$, ce qui nous montre que l'erreur est donnée par

$$e_n(u)_V^2 = \sum_{k>n} (c_k^*)^2. \tag{2.57}$$

Cette série converge puisqu'on sait que les c_k^* sont de carré sommable avec

$$\sum_{k\geq 1} (c_k^*)^2 = \sum_{k\geq 1} |c_k|^2 = \|u\|_V^2, \tag{2.58}$$

grace à la propriété de permutation. Plus généralement, en notant $\mathbf{c} = (c_k)_{k\geq 1}$ et $\mathbf{c}^* = (c_k^*)_{k\geq 1}$, on a

$$\|\mathbf{c}^*\|_{\ell^p} = \|\mathbf{c}\|_{\ell^p}, \tag{2.59}$$

pour tout p. Supposons à présent que la suite $\mathbf{c} = (c_k)_{k\geq 1}$ appartienne aussi à $\ell^p(\mathbb{N})$ pour une valeur $p < 2$. On peut alors écrire

$$e_n(u)_V^2 = \sum_{k>n} (c_k^*)^{2-p} (c_k^*)^p \leq (c_{n+1}^*)^{2-p} \sum_{k>n} (c_k^*)^p \leq C^p (c_{n+1}^*)^{2-p}, \tag{2.60}$$

avec $C = \|\mathbf{c}\|_{\ell^p} = \|\mathbf{c}^*\|_{\ell^p}$, où on a utilisé la décroissance des c_k^*. On a d'autre part

$$(n+1)(c_{n+1}^*)^p \leq \sum_{k=1}^{n+1} (c_k^*)^p \leq C^p. \tag{2.61}$$

En combinant ces deux estimations, on obtient

$$e_n(u)_V^2 \leq C^2 (n+1)^{1-\frac{2}{p}}, \tag{2.62}$$

soit

$$e_n(u)_V \leq C(n+1)^{-s} \leq Cn^{-s}, \quad s := \frac{1}{p} - \frac{1}{2}. \tag{2.63}$$

En conclusion nous venons de montrer que la propriété de sommabilité ℓ^p des coefficients de u entraîne la décroissance en n^{-s} de l'erreur d'approximation non-linéaire, avec $s := \frac{1}{p} - \frac{1}{2}$. Là aussi on a pas une implication réciproque mais presque : on peut montrer (**exercice**) que si $e_n(u)_V \leq Cn^{-s}$ pour tout $n \geq 1$ alors $\mathbf{c} \in \ell^q$ pour tout $q > p$. On a ainsi obtenu le résultat suivant.

Théorème 2.5 *Pour l'approximation à n termes dans une base orthonormée d'un espace de Hilbert V, on a pour tout $s > 0$ et $\varepsilon > 0$,*

$$\mathbf{c} \in \ell^p \implies u \in \mathcal{A}^s \implies \mathbf{c} \in \ell^{p+\varepsilon}, \tag{2.64}$$

avec $\frac{1}{p} = \frac{1}{2} + s$ et $\mathbf{c} = (c_k)_{k\geq 1}$ la suite des coordonnées de u.

Remarque 2.5 *Il est en fait possible de caractériser exactement l'espace $\mathcal{A}^s$ par une propriété de sommabilité qui fait intervenir une variante de l'espace ℓ^p. On dit qu'une*

suite $(c_k)_{k\geq 1}$ est faiblement ℓ^p-sommable, ou dans l'espace $\ell^p_w(\mathbb{N})$, si et seulement si il existe une constante C telle que

$$c_k^* \leq Ck^{-1/p}, \quad k \geq 1. \tag{2.65}$$

On voit aisément que cette propriété est impliquée par $\mathbf{c} \in \ell^p$ puisque

$$k(c_k^*)^p \leq \sum_{j=1}^{k} (c_j^*)^p \leq \|\mathbf{c}\|_{\ell^p}^p, \tag{2.66}$$

*et on a ainsi $\ell^p \subset \ell^p_w$ mais cette inclusion est stricte. Par exemple la suite $(1/k)_{k\geq 1}$ appartient à ℓ^1_w mais pas à ℓ^1. Il est aussi facile de vérifier que $\ell^p_w \subset \ell^q$ pour tout $q > p$. On montre (***exercice***) que l'espace $\mathcal{A}^s$ est alors exactement caractérisé par*

$$\mathbf{c} \in \ell^p_w \iff u \in \mathcal{A}^s, \tag{2.67}$$

avec $\frac{1}{p} = \frac{1}{2} + s$ et $\mathbf{c} = (c_k)_{k\geq 1}$ la suite des coordonnées de u.

En résumé la vitesse n^{-s} pour l'approximation non-linéaire d'une fonction u dans une base orthonormée est gouvernée par la sommabilité ℓ^p de la suite $\mathbf{c}$ de ses coefficients pour $p < 2$ tel que $\frac{1}{p} = \frac{1}{2} + s$. Noter que plus p est petit, plus la valeur de s est grande. En ce sens l'exposant de sommabilité ℓ^p exprime une propriété de concentration de la suite $\mathbf{c}$ qui se traduit par le fait qu'elle est bien approchée dans ℓ^2 par ses n plus grand termes. De même qu'on a introduit les espaces V^s pour l'analyse de l'approximation linéaire, on peut introduire pour $p < 2$ des espaces

$$W^p := \{u \in V \,:\, \mathbf{c} \in \ell^p(\mathbb{N})\}, \tag{2.68}$$

et les munir de la norme $\|u\|_{W^p} := \|\mathbf{c}\|_{\ell^p}$. On a ainsi $W^p \subset \mathcal{A}^s \subset W^{p+\varepsilon}$ pour $s = \frac{1}{p} - \frac{1}{2}$.

Dans le cas particulier de la base des séries de Fourier, l'espace W^p, ne peut pas être identifié avec un espace de régularité classique, tel que l'espaces de Sobolev périodiques H^s_{per} qu'on a identifié à V^s pour l'approximation linéaire. Les fonctions de W^p n'ont pas de dérivées dans L^2 et l'injection de W^p dans L^2 n'est pas compacte, tout simplement parce que l'injection de ℓ^p dans ℓ^2 n'est pas compacte (**exercice**). Dans le cas de bases d'ondelettes sur un domaine $D \subset \mathbb{R}^d$, l'espace W^p s'identifie avec l'espace de Besov $B^s_{p,p}$ avec $\frac{1}{p} = \frac{1}{2} + \frac{s}{d}$. Ceci nous montre que s dérivées dans L^p assurent la vitesse d'approximation à n termes $\mathcal{O}(n^{-s/d})$, ce qui est à rapprocher du résultat général indiqué dans la remarque 2.1 pour l'approximation par des fonctions polynomiales par morceaux sur des partitions adaptatives.

Noter que lorsque s devient grand, on a $p < 1$ ce qui signifie que l'espace ℓ^p n'est plus un espace de Banach mais un espace de quasi-Banach (l'inégalité triangulaire est vérifiée à une constant multiplicative près). Noter aussi que quand $p \to 0$, la quantité $\|\mathbf{c}\|_{\ell^p}^p = \sum_{k\geq 1} |c_k|^p$ tend vers la taille du support de la suite, que l'on note parfois

$$\|\mathbf{c}\|_{\ell^0} = \#\{k \,:\, c_k \neq 0\}, \tag{2.69}$$

même si il ne s'agit pas d'une norme.

L'approximation à n termes de $u \in V$ dans une base orthonormée d'un espace de Hilbert V est équivalente à celle de sa suite de coefficients $\mathbf{c} \in \ell^2$ par l'ensemble V_n des suites ayant au plus n-termes non-nuls, c'est à dire

$$V_n = \{\mathbf{d} \ : \ \|\mathbf{d}\|_{\ell^0} \leq n\}, \tag{2.70}$$

On a en effet $e_n(u)_V = e_n(\mathbf{c})_{\ell^2}$. Plus généralement, si on considère l'approximation de suites $\mathbf{c} \in \ell^q$ par de telles suites, on voit que la meilleure approximation est à nouveau obtenue en conservant les n plus grands coefficients et en mettant à zéro les autres.

$$e_n(\mathbf{c})_{\ell^q}^q = \sum_{k \notin E_n} |c_k|^q = \sum_{k>n} (c_k^*)^q. \tag{2.71}$$

Par les mêmes techniques qu'on a utilisé dans le cas $q = 2$ (**exercice**), on obtient un résultat plus général (parfois appelé lemme de Stechkin) dont on fera usage à la fin du cours.

Théorème 2.6 *Si $\mathbf{c} \in \ell^p$ avec $p < q$, l'erreur d'approximation à n termes vérifie*

$$e_n(\mathbf{c})_{\ell^q} \leq C n^{-s}, \tag{2.72}$$

où $s = \frac{1}{p} - \frac{1}{q}$ et $C = \|\mathbf{c}\|_{\ell^p}$.

2.3 Optimalité : épaisseurs et entropies

Dans la section précèdente on a examiné certaines méthodes d'approximation, linéaire et non-linéaire, et on a caractérisé - en terme de régularité ou de sommabilité - les espaces de fonctions u qui assurent une vitesse de convergence $e_n(u)_V = \mathcal{O}(n^{-s})$ pour ces méthodes. Par exemple, on a identifié les espaces de Hölder $\mathcal{C}^s([0,1])$ pour l'approximation par les constantes par morceaux en norme L^∞, l'espace $H^s_{per}([0,1])$ pour l'approximation par des séries de Fourier, l'espace W^p des fonctions à coefficients ℓ^p sommable avec $\frac{1}{p} = \frac{1}{2} + s$ pour l'approximation à n termes... Si $\mathcal{K}$ représente à chaque fois la boule unité de ces espaces, on obtient donc pour la méthode considérée une borne

$$e_n(\mathcal{K})_V = \sup_{u \in \mathcal{K}} e_n(u)_V \leq C n^{-s}. \tag{2.73}$$

pour l'erreur maximale sur $\mathcal{K}$. C'est aussi le cas plus généralement si $\mathcal{K}$ est n'importe quelle partie bornée des ces espaces.

A l'inverse, partons d'une classe $\mathcal{K} \subset V$ de fonctions données qui nous intéresse, par exemple une classe à laquelle on sait qu'appartient la fonction u qu'on veut reconstruire à partir de données ponctuelles. Il est naturel de se poser une toute autre question : pour cette classe y-a-t-il une méthode d'approximation qui est en un certain sens optimale, en particulier du point de vue de sa vitesse de convergence. Par exemple pour la boule unité de $H^s_{per}([0,1]^d)$, la vitesse $n^{-s/d}$ obtenue avec les séries de Fourier aurait elle pu être améliorée par une autre méthode d'approximation ?

Cette question est bien comprise dans le cas de l'approximation linéaire. Pour une classe $\mathcal{K} \subset V$ que l'on suppose compacte pour la norme $\|\cdot\|_V$ et pour $n \geq 0$, on définit la *n-épaisseur de Kolmogorov* (n-width) par

$$d_n(\mathcal{K})_V := \inf_{\dim(V_n)=n} \max_{u\in\mathcal{K}} \min_{v\in V_n} \|u-v\|_V = \inf_{\dim(V_n)=n} \mathrm{dist}(\mathcal{K}, V_n). \tag{2.74}$$

Autrement dit on met en compétition tous les espaces de dimension n pour approcher au mieux uniformément la classe $\mathcal{K}$. Noter que dans la définition, on peut remplacer $\dim(V_n) = n$ par $\dim(V_n) \leq n$. Dans cette formule, le maximum sur u et le minimum sur v sont atteints, mais il n'est pas toujours vrai que l'infimum sur V_n est atteint par un espace $\overline{V}_n$ qui pourrait alors être qualifié d'optimal. Sinon, on se contente de parler d'espaces "presque optimaux", par exemple qui atteignent l'infimum à $\varepsilon > 0$ près ou à une constante multiplicative près. On voit que la suite des épaisseurs $(d_n(\mathcal{K})_V)_{n\geq 0}$ est décroissante.

Notons que si V est un espace de Hilbert on peut écrire

$$d_n(\mathcal{K})_V := \inf_{\dim(V_n)=n} \max_{u\in\mathcal{K}} \|u - P_{V_n} u\|_V, \tag{2.75}$$

où P_{V_n} est la projection orthogonale. Dans ce cas, il convient de noter les analogies et différences avec *l'analyse en composante principale* (ACP) d'une variable aléatoire vectorielle $\mathbf{x}$ centrée dans un espace de Hilbert V de dimension finie ou infinie, qui permet de définir les espaces $\overline{V}_n$ optimaux pour le problème

$$\sigma_n(\mathbf{x})_V^2 := \min_{\dim(V_n)=n} \mathbb{E}(\|\mathbf{x} - P_{V_n}\mathbf{x}\|_V^2). \tag{2.76}$$

Dans cette expression, l'erreur est mesurée dans un sens quadratique moyen, alors qu'on la mesure uniformément sur $\mathcal{K}$ dans la définition de $d_n(\mathcal{K})_V$. En particulier, si la variable X est supportée dans la classe $\mathcal{K}$ on a toujours $\sigma_n(\mathbf{x})_V \leq d_n(\mathcal{K})_V$ (**exercice**).

Les épaisseurs de Kolmogorov $(d_n(\mathcal{K}))_{n\geq 0}$ nous indiquent donc ce qu'on peut attendre de mieux si on utilise une méthode d'approximation linéaire et qu'on cherche une borne d'erreur garantie uniformément sur l'ensemble de la classe $\mathcal{K}$. La recherche pratique d'espaces $\overline{V}_n$ optimaux ou presque optimaux dans la définition de $d_n(\mathcal{K})_V$ est un problème difficile qui sera en partie résolu dans le chapitre 4 par la méthode des bases réduites. Dans cette section nous allons discuter de techniques générales qui permettent d'estimer $d_n(\mathcal{K})_V$, et en particulier sa décroissance avec n.

Donnons auparavant quelques variantes à la définition des épaisseurs de Kolmogorov. On appelle *nombres d'approximation* les quantités

$$a_n(\mathcal{K})_V := \inf_{\mathrm{rang(L_n)}\leq \mathrm{n}} \max_{u\in\mathcal{K}} \|u - L_n u\|_V, \tag{2.77}$$

où l'infimum est pris sur toutes les applications L_n linéaires de rang inférieur ou égal à n (c'est à dire dont l'image est de dimension au plus n). On a naturellement

$$d_n(\mathcal{K})_V \leq a_n(\mathcal{K})_V, \tag{2.78}$$

et cette inégalité devient une égalité dans le cas particulier où V est un espace de Hilbert, puisque P_{V_n} est linéaire et donne la meilleure approximation. L'inégalité est en générale

stricte dans un espace de Banach général puisqu'on a remarqué que la meilleure approximation dans un espace V_n n'est pas obtenue par une projection linéaire. Le théorème de Kadec-Snobar déjà cité permet au mieux d'affirmer que

$$a_n(\mathcal{K})_V \leq (1+\sqrt{n})d_n(\mathcal{K})_V. \tag{2.79}$$

On peut être moins exigeant et demander uniquement que l'approximation soit réalisée par une application continue, ce qui revient à définir la quantité

$$\bar{d}_n(\mathcal{K})_V := \inf_{\mathrm{rang(F_n)}\leq \mathrm{n}} \max_{u\in\mathcal{K}} \|u - F_n(u)\|_V, \tag{2.80}$$

où l'infimum est ici pris sur toutes les applications F_n continues (mais pas forcément linéaires) dont l'image est contenue dans un espace vectoriel de dimension inférieur à n. Noter que l'opérateur de meilleure approximation sur V_n peut être discontinu et que l'inégalité $d_n(\mathcal{K})_V \leq \bar{d}_n(\mathcal{K})_V$ pourrait donc être stricte. On peut cependant montrer (**exercice difficile**) que l'on a

$$\bar{d}_n(\mathcal{K})_V = d_n(\mathcal{K})_V. \tag{2.81}$$

Indication : pour $\varepsilon > 0$ et V_n de dimension n arbitrairement fixés, utiliser un recouvrement de $\mathcal{K}$ par des boules suffisament petites et une partition de l'unité pour construire $F_n : V \to V_n$ continue telle que $\|u - F_n(u)\|_V \leq \min_{v\in V_n} \|u - v\|_V + \varepsilon$ pour tout $u \in \mathcal{K}$.

Dans la section précèdente on a mis en évidence des classes $\mathcal{K}$ et des approximations par des familles linéaires $(V_n)_{n\geq 0}$ spécifiques pour lesquelles on a

$$e_n(\mathcal{K})_V := \max_{u\in\mathcal{K}} \min_{v\in V_n} \|u - v\|_V \leq Cn^{-s}. \tag{2.82}$$

Ceci donne immédiatement des bornes supérieures sur les épaisseurs de Kolmogorov de $\mathcal{K}$ suivant

$$d_n(\mathcal{K})_V \leq Cn^{-s}. \tag{2.83}$$

Afin de savoir si on a fait avec ces $(V_n)_{n\geq 0}$ un choix optimal parmi toutes les familles linéaires, il nous faudrait établir une inégalité dans le sens inverse

$$d_n(\mathcal{K})_V \geq cn^{-s}, \tag{2.84}$$

ce qui nécessite d'établir des bornes inférieures sur les $d_n(\mathcal{K})_V$. Notons que $d_n(\mathcal{K})_V \geq \delta$ signifie que, pour tout espace V_n de dimension n, il existe une fonction $u \in \mathcal{K}$ telle que $\|u - v\|_V \geq \delta$ pour tout $v \in V_n$.

Etudions un exemple simple : on considère la classe donnée par la boule unité des fonctions lipschitziennes sur l'intervalle $D = [0, 1]$,

$$\mathcal{K} := \{u \ : \ \|u\|_{\mathrm{Lip}} := \max\{\|u\|_{L^\infty}, \|u'\|_{L^\infty}\} \leq 1\}. \tag{2.85}$$

On a vu que l'approximation par les fonctions constantes par morceaux sur des partitions uniformes donne l'estimation $e_n(u)_{L^\infty} \leq \frac{1}{2}\|u'\|_{L^\infty} n^{-1}$ et par conséquent

$$d_n(\mathcal{K})_{L^\infty} \leq \frac{1}{2} n^{-1}. \tag{2.86}$$

Il est possible d'obtenir une estimation similaire en utilisant des espaces V_n de fonctions continues (**exercice** : utiliser des fonctions affines par morceaux pour un maillage bien choisi).

Afin d'obtenir une borne inférieure, on considère les $n+1$ points

$$x^i = i/n, \quad i = 0, \dots, n. \tag{2.87}$$

Soit V_n n'importe quel espace de dimension n de fonctions continues. On lui associe l'espace

$$W_n := \{(v(x^0), \dots, v(x^n)) \,:\, v \in V_n\} \subset \mathbb{R}^{n+1}, \tag{2.88}$$

qui est au plus de dimension n. Il existe donc un vecteur non nul $\mathbf{b} = (b_0, \dots, b_n) \in \mathbb{R}^{n+1}$ qui est orthogonal à W_n, c'est à dire tel que

$$b_0 v(x^0) + \cdots + b_n v(x^n) = 0, \quad v \in V_n. \tag{2.89}$$

Pour chaque j on pose

$$\varepsilon_j = \operatorname{sign}(b_j), \tag{2.90}$$

c'est à dire 1 si $b_j > 0$ et -1 si $b_j < 0$ et 0 si $b_j = 0$, et on définit une fonction u continue, et affine sur chaque $[x^j, x^{j+1}]$, telle que

$$u(x^j) = \frac{1}{2n}\varepsilon_j, \quad j = 0, \dots, n. \tag{2.91}$$

Par construction les fonctions u et u' sont inférieures à 1 en valeurs absolues, ce qui nous montre que $u \in \mathcal{K}$. La relation d'orthogonalité (2.89) nous indique que pour tout $v \in V_n$, il existe au moins une valeur de $j = j(v)$ telle que $\operatorname{sign}(b_j) \neq \operatorname{sign}(v(x^j))$, et par conséquent

$$\|u - v\|_{L^\infty} \geq |u(x^j) - v(x^j)| \geq \frac{1}{2n}. \tag{2.92}$$

On a ainsi établi une borne inférieure sur $d_n(\mathcal{K})_{L^\infty}$ qui est dans ce cas précisément égale à la borne supérieure : on a donc exactement

$$d_n(\mathcal{K})_{L^\infty} = \frac{1}{2} n^{-1}. \tag{2.93}$$

Ce principe de minoration par le bas peut se généraliser (**exercice**) pour un domaine borné $D \subset \mathbb{R}^d$ par exemple $D = [0,1]^d$: pour la boule unité

$$\mathcal{K} := \{u \,:\, \|u\|_{\mathrm{Lip}} := \max\{\|u\|_{L^\infty}, \|\nabla u\|_{L^\infty}\} \leq 1\}. \tag{2.94}$$

on obtient ainsi

$$c n^{-1/d} \leq d_n(\mathcal{K})_{L^\infty} \leq C n^{-1/d}. \tag{2.95}$$

Ceci nous montre que la malédiction des grandes dimensions frappe toutes les méthodes d'approximation linéaires, pas seulement les fonctions constantes par morceaux, lorsqu'on considère la classe des fonctions lipschitziennes. Intuitivement, lorsque le nombre d de

variables augmente, cette classe devient beaucoup trop large pour être uniformément bien approchée par des espaces de dimension n petite. On montre de même que

$$d_n(\mathcal{K})_{L^\infty} \sim n^{-s/d}, \tag{2.96}$$

lorsque $\mathcal{K}$ est la boule unité de $\mathcal{C}^s(D)$.

La technique que nous venons d'utiliser est très spécifique à la mesure d'erreur en norme L^∞. Nous allons donner deux approches plus générales permettant d'établir des bornes inférieures pour les épaisseurs de Kolmogorov. La première approche découle du théorème général suivant.

Théorème 2.7 *Soit V un espace de Banach, $W \subset V$ un sous espace de dimension $n+1$, et*

$$B_W := \{u \in W \ : \ \|u\|_V \leq 1\}, \tag{2.97}$$

sa boule unité. Alors $d_n(B_W)_V = 1$.

Noter que l'inégalité $d_n(B_W)_V \leq 1$ est évidente puisque tous les éléments de B_W sont à distance au plus 1 de l'origine et le résultat principal est ici que $d_n(B_W)_V \geq 1$. Ce résultat est assez intuitif et sa preuve est évidente dans le cas où V est un espace de Hilbert : pour tout espace V_n de dimension n on remarque qu'il existe un vecteur non-nul $w \in W$ qui est orthogonal à V_n. Après renormalisation, $u = \|w\|^{-1}w$ appartient à B_W et on a ainsi

$$\|u - P_{V_n}u\|_V = \|u\|_V = 1, \tag{2.98}$$

ce qui entraine $\mathrm{dist}(B_W, V_n) \geq 1$, et par conséquent $d_n(B_W)_V \geq 1$. Pour un espace de Banach général, la preuve est beaucoup moins évidente et découle d'un résultat profond de topologie admis ici : le théorème d'antipodalité de Borsuk-Ulham.

Théorème 2.8 *Soit F une application continue d'une sphére unité S_n de dimension n (c'est à dire la frontière d'une boule unité d'un espace de dimension $n+1$) vers un espace vectoriel de dimension n. Alors il existe $x \in S_n$ tel que $F(x) = F(-x)$.*

On pourra démontrer (**exercice**) ce résultat dans le cas simple $n = 1$ c'est à dire pour une application continue allant du cercle S_1 vers $\mathbb{R}$, c'est à dire une fonction continue périodique. Noter que la norme utilisée pour définir la sphére unité ne joue aucun rôle ici : on peut se ramener à la sphère euclidienne par une transformation continue qui préserve l'antipodalité (**exercice**).

Le Théorème 2.7 se déduit alors par l'argument suivant : en notant $\delta = d_n(B_W)_V$ on a vu que l'on a aussi $\delta = \bar{d}_n(B_W)$, ce qui signifie que pour tout $\varepsilon > 0$ existe une application F continue et dont l'image est contenue un espace vectoriel V_n de dimension n telle que

$$\|u - F(u)\|_V \leq \delta + \varepsilon, \quad u \in B_W. \tag{2.99}$$

La frontière $S = \partial B_W$ de B_W étant une sphère de dimension n, le théorème de Borsuk-Ulham entraîne l'existence d'un $u^* \in \partial B$ tel que $F(u^*) = F(-u^*)$. Ceci entraine que

$$2(\delta + \varepsilon) \geq \|u^* - F(u^*)\|_V + \|F(-u^*) - (-u^*)\|_V \geq \|2u^*\|_V = 2, \tag{2.100}$$

c'est à dire $\delta + \varepsilon \geq 1$ pour tout $\varepsilon > 0$, et donc $\delta \geq 1$, ce qui prouve le Théorème 2.7.

Par changement d'échelle, ce théorème implique que si $B_W(r)$ est une boule de dimension $n+1$ et de rayon $r > 0$, on a $d_n(B_W(r))_V \geq r$. Plus généralement, si $\mathcal{K} \subset V$ est un ensemble qui contient une boule $B_W(r)$ de dimension $n+1$ et de rayon $r > 0$, on a $d_n(\mathcal{K})_V \geq r$. On définit la *n-épaisseur de Bernstein* de $\mathcal{K}$ par

$$b_n(\mathcal{K})_V := \max\{r > 0 \ : \ \exists W, \ \dim(W) = n+1, \ B_W(r) \subset \mathcal{K}\} \tag{2.101}$$

c'est à dire le rayon maximal d'une boule de dimension $n+1$ centrée à l'origine qu'on peut inclure dans $\mathcal{K}$. On a par conséquent la borne inférieure générale

$$d_n(\mathcal{K})_V \geq b_n(\mathcal{K})_V, \quad n \geq 0. \tag{2.102}$$

On peut (**exercice**) utiliser cette borne inférieure pour retrouver le résultat $d_n(\mathcal{K})_V \geq cn^{-1}$ pour $V = L^\infty([0,1])$ et $\mathcal{K}$ la boule unité de fonction lipschitziennes en montrant que $b_n(\mathcal{K})_V \geq cn^{-1}$. On peut aussi l'utiliser pour d'autres normes : considérer par exemple $V = L^2([0,1])$ et $\mathcal{K}$ la boule unité $H^s_{per}([0,1])$ et montrer (**exercice**) que $b_n(\mathcal{K})_V \geq cn^{-s}$. Cette approche fonctionne aussi en dimension $d > 1$ pour établir des bornes inférieures de la forme $cn^{-s/d}$.

Une autre approche pour minorer les épaisseurs de Kolmogorov est de quantifier la taille de $\mathcal{K}$ de la manière suivante : pour $\varepsilon > 0$ on définit le nombre de recouvrement (covering number) $N_\varepsilon = N_\varepsilon(\mathcal{K})_V$ comme la plus petite valeur de N telle qu'il existe un recouvrement de $\mathcal{K}$ par N boules de rayon ε : il existe $v^1, \ldots, v^N \in V$ tels que

$$\mathcal{K} \subset \bigcup_{i=1}^N B(v^i, \varepsilon), \quad B(v^i, \varepsilon) := \{v \in V \ : \ \|v - v^i\|_V \leq \varepsilon\}. \tag{2.103}$$

De tels recouvrements finis existent pour tout $\varepsilon > 0$ puisque $\mathcal{K}$ est compact, et on a $N_\varepsilon = 1$ pour ε suffisament grand. Une notion proche est le nombre d'empilement (packing number) $M_\varepsilon = M_\varepsilon(\mathcal{K})_V$ qui est défini comme la plus grande valeur de M telle qu'il existe $u^1, \ldots, u^M \in \mathcal{K}$ ayant la propriété d'écartement

$$\|u^i - u^j\|_V > \varepsilon, \quad i \neq j. \tag{2.104}$$

Ces quantités sont presque équivalentes au sens où

$$M_{2\varepsilon} \leq N_\varepsilon \leq M_\varepsilon, \tag{2.105}$$

En effet si $u^1, \ldots, u^{M_\varepsilon} \in \mathcal{K}$ est un empilement maximal pour l'écartement ε, alors tout $u \in \mathcal{K}$ doit être à distance inférieure ou égale à ε d'au moins un u^i (sinon on pourrait l'ajoûter aux u^i ce qui contredirait la maximalité de M_ε). Cela signifie que $\mathcal{K}$ est recouvert par les $B(u^i, \varepsilon)$ et entraine donc l'inégalité $N_\varepsilon \leq M_\varepsilon$. D'autre part si $u^1, \ldots, u^{M_{2\varepsilon}} \in \mathcal{K}$ sont distants deux à deux de plus que 2ε alors toute boule de rayon ε ne peut contenir qu'un seul de ces points, ce qui entraîne l'inégalité $M_{2\varepsilon} \leq N_\varepsilon$.

On définit les *entropies métrique de Kolmogorov* comme les quantités

$$H_\varepsilon = H_\varepsilon(\mathcal{K})_V := \lceil \log_2(N_e) \rceil, \tag{2.106}$$

c'est à dire le plus petit entier $n = n(\varepsilon)$ tel que $\mathcal{K}$ puisse être recouvert par 2^n boules de taille ε. Cette notion est à relier à celle de compression de l'information ou codage *avec perte* : on peut décider de coder $u \in \mathcal{K}$ par le symbole i de la boule auquel il appartient, et cette information permet de reconstruire u^i à la place de u en faisant une erreur inférieure à ε. Comme il y a $N_\varepsilon \leq 2^{H_\varepsilon}$ symboles, H_ε peut être vu comme le nombre minimal de bits $\{0,1\}$ nécessaire pour coder tout élément de $\mathcal{K}$ avec une perte de précision ε. Une notion inverse est celle de nombres d'entropies donnée par

$$\varepsilon_n = \varepsilon_n(\mathcal{K})_V := \min\{\varepsilon > 0 \ : \ H_\varepsilon \leq n\}, \tag{2.107}$$

qui est donc la précision de codage qu'on peut atteindre avec au plus n bits.

Remarque 2.6 *Il ne faut pas faire de confusion avec l'entropie de Shannon $H = -\sum p_i \log_2(p_i)$ qui en théorie de l'information quantifie le nombre de bits minimal nécessaire en moyenne pour le codage sans perte d'un nombre fini de symboles ayant chacun une certaine probabilité p_i d'apparition.*

Il est facile d'estimer N_ε lorsque $\mathcal{K} = B$ est la boule unité d'un sous-espace de dimension finie n. On suppose $\varepsilon < 1$, sinon on a trivialement $N_\varepsilon = 1$, et par changement d'échelle on remarque que le volume d'une boule de rayon ε est égal à ε^n fois celui de B. Il vient par la propriété de recouvrement que

$$|B| \leq N_\varepsilon \varepsilon^n |B|, \tag{2.108}$$

Inversement, on se donne un empilement maximal $u^1, \ldots, u^{M_\varepsilon} \in B$ ce qui signifie que les boules $B(u^i, \varepsilon/2)$ sont deux à deux disjointes, et contenues dans la boule $(1+\varepsilon/2)B$. Par conséquent

$$M_\varepsilon(\varepsilon/2)^n |B| \ \leq (1+\varepsilon/2)^n |B| \leq (3/2)^n |B|. \tag{2.109}$$

On a ainsi établi l'encadrement

$$\varepsilon^{-n} \leq N_\varepsilon \leq (\varepsilon/3)^{-n}. \tag{2.110}$$

Nous allons maintenant relier les d'entropies avec les épaisseurs. Pour $n \geq 0$, prenons $\delta > d_n(\mathcal{K})_V$, et $R > \delta$ tel que $\mathcal{K} \subset B(0,R)$. Il existe un espace $V_n \subset V$ de dimension n tel que tout $u \in \mathcal{K}$ est à distance δ de V_n. Comme $\|u\|_V \leq R$, on voit aussi que tout $u \in \mathcal{K}$ est à distance δ d'un point de la boule B de V_n centrée en 0 et de rayon $R+\delta$. La boule B peut être recouverte par N boules de la forme $B(u^i, \delta)$ avec $N \leq (R+\delta)^n(\delta/3)^{-n} \leq (2R)^n(\delta/3)^{-n}$, et par conséquent les N boules $B(u^i, 2\delta)$ recouvrent $\mathcal{K}$. On a ainsi établi l'inégalité

$$N_{2\delta}(\mathcal{K})_V \leq \Big(\frac{6R}{\delta}\Big)^n. \tag{2.111}$$

Si on suppose à présent que

$$d_n(\mathcal{K})_V \leq Cn^{-s}, \quad n \geq 1, \tag{2.112}$$

alors pour tout $\varepsilon \geq 2Cn^{-s}$ on trouve (quitte à modifier à chaque fois la constante C) que $N_\varepsilon \leq (Cn^s)^n$ soit

$$H_\varepsilon \leq \lceil \log N_{2\delta}(\mathcal{K})_V \rceil \leq Cn \log_2(n^s) \leq C\varepsilon^{-1/s} \log_2(1/\varepsilon). \tag{2.113}$$

Au facteur logarithmique $\log_2(1/\varepsilon)$ près, on obtient donc $H_\varepsilon \leq C\varepsilon^{-1/s}$ ce qui signifie que les nombres d'entropie décroissent aussi suivant

$$\varepsilon_n(\mathcal{K})_V \leq Cn^{-s}. \tag{2.114}$$

Ce résultat peut être prouvé rigoureusement par une analyse plus fine : *l'inégalité de Carl* affirme que pour tout $s > 0$, il existe une constante C_s telle que

$$\sup_{n\geq 0}(n+1)^s\varepsilon_n(\mathcal{K})_V \leq C_s \sup_{n\geq 0}(n+1)^s d_n(\mathcal{K})_V. \tag{2.115}$$

On peut en particulier cette inégalité pour montrer que les épaisseurs de Kolmogorov d'une classe ne décroissent pas plus vite que n^{-s} si on sait qu'il en est ainsi pour les nombres d'entropie.

Si l'on s'intéresse à présent à l'optimalité de méthodes non-linéaires sur des classes de fonctions, on peut chercher à introduire des notions d'épaisseurs non-linéaires. La première idée qui vient à l'esprit est de remplacer les espaces de dimension n par des variétés de dimension n dans la définition de d_n. Par exemple, on peut introduire la classe $\mathcal{M}_n$ de toutes les ensembles décrits par un paramétrage régulier à n variables, c'est à dire de la forme $S = \{g(x) \ : \ x \in P\}$ où $P \subset \mathbb{R}^n$ et $g : \mathbb{R}^n \to V$ est $\mathcal{C}^\infty$, et proposer la quantité

$$\inf_{S\in\mathcal{M}_n} \max_{u\in\mathcal{K}} \min_{v\in S} \|u - v\|_V. \tag{2.116}$$

Cette approche n'est pas la bonne, car même pour $n = 1$ on peut trouver (**exercice difficile**) une suite de courbes régulières $(S^j)_{j\geq 1}$ qui est dense dans $\mathcal{K}$ au sens où $\max_{u\in\mathcal{K}} \min_{v\in S^j} \|u - v\|_V \to 0$ quand $j \to \infty$. De ce fait la quantité ci-dessus est nulle pour tout compact $\mathcal{K}$.

Afin d'éviter ces familles d'approximation pathologiques, on adopte un point de vue un peu plus restrictif en imposant une forme de stabilité sur l'application qui associe à une fonction u les n paramètres décrivant son approximation. On définit ainsi la n-épaisseur non-linéaire par

$$\delta_n(\mathcal{K})_V := \inf_{E,D} \max_{u\in\mathcal{K}} \|u - D(E(u))\|_V, \tag{2.117}$$

où l'infimum est pris sur tous les schémas "encodage-décodage" de la forme

$$E : \mathcal{K} \to \mathbb{R}^n \quad \text{et} \quad D : \mathbb{R}^n \to V, \tag{2.118}$$

où E et D sont supposées *continues*. On note que $d_n(\mathcal{K})_V = \bar{d}_n(\mathcal{K})_V$ peut-être défini par la même formule mais en restreignant D aux applications linéaires de $\mathbb{R}^n$ dans V, ce qui montre que $\delta_n(\mathcal{K})_V \leq d_n(\mathcal{K})_V$.

Si $\mathcal{K} = B_W$ est la boule unité d'un espace $W \subset V$ de dimension $n+1$, on peut raisonner comme on l'a fait pour minorer $\bar{d}_n(B_W)_V$: en posant $\delta = \delta_n(B_W)_V$, et pour tout $\varepsilon > 0$, il existe $E : B_W \to \mathbb{R}^n$ et $D : \mathbb{R}^n \to V$ continues tels que

$$\|u - D(E(u))\|_V \leq \delta + \varepsilon, \quad u \in B_W, \tag{2.119}$$

et par ailleurs le théorème de Borsuk-Uhlam indique qu'il existe $u^* \in \partial B_W$ tel que $E(u^*) = E(-u^*)$, ce qui entraine

$$2(\delta+\varepsilon) \geq \|u^* - D(E(u^*))\|_V + \| - u^* - D(E(-u^*))\|_V \geq \|2u^*\|_V = 2, \tag{2.120}$$

et par conséquent $\delta_n(B_W)_V \geq 1$. On obtient de cette manière pour tout compact $\mathcal{K}$ une minoration de l'épaisseur non-linéaire par l'épaisseur de Bernstein

$$\delta_n(\mathcal{K})_V \geq b_n(\mathcal{K})_V, \tag{2.121}$$

similaire à celle obtenue pour l'épaisseur linéaire.

Ce résultat permet de montrer en particulier que si $\mathcal{K}$ est la boule unité des fonctions $\mathcal{C}^s([0,1]^d)$ on aura comme pour l'épaisseur linéaire $d_n(\mathcal{K})_{L^\infty}$ la borne inférieure

$$\delta_n(\mathcal{K})_{L^\infty} \geq cn^{-s/d}, \quad n \geq 1, \tag{2.122}$$

autrement dit l'approximation non-linéaire ne permet pas de contourner la malédiction des grandes dimension.

Remarque 2.7 *La vitesse $n^{-s/d}$ fait apparaître une compensation entre la régularité et le nombre de variables, ce qui peut suggérer qu'une manière de contourner la malédiction des grandes dimension est de considérer des classes de fonctions $\mathcal{C}^\infty$. Ce n'est en fait pas généralement suffisant, comme le montre par exemple le résultat suivant dû à Novak-Wozniakovski : pour la classe*

$$\mathcal{K} := \mathcal{K}_d = \{u \in \mathcal{C}^\infty([0,1]^d) \ : \ \|\partial^\alpha u\|_{L^\infty} \leq 1, \quad \alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d\}, \tag{2.123}$$

on a

$$\min\left\{n \geq 0 \ : \ d_n(\mathcal{K}_d)_{L^\infty} \geq \alpha\right\} \ \geq \exp(\beta d), \quad d \geq 1 \tag{2.124}$$

où α et β sont des nombres strictement positifs fixés. On verra dans le dernier chapitre que la malédiction des grandes dimensions peut-être contourné pour des fonctions qui ont certaines propriétés d'anisotropie (au sens où certaines variables sont plus "actives" que d'autres dans les variations de la fonction).

Pour terminer ce tour d'horizon des quantités qui mesurent d'optimalité des approximation sur une classe $\mathcal{K}$, on peut chercher à quantifier la possibilité de reconstruire au mieux les fonctions $\mathcal{K}$ à partir d'un nombre n de mesures linéaires. Si $\lambda_1, \ldots, \lambda_n \in V^*$ sont des formes linéaires continues, la quantité

$$\inf_R \sup_{u \in \mathcal{K}} \|u - R(\lambda_1(u), \ldots, \lambda_n(u))\|_V \tag{2.125}$$

où l'infimum est pris sur toute les applications $R : \mathbb{R}^n \to V$, mesure avec quelle erreur maximale les mesures $\lambda_1(u), \ldots, \lambda_n(u)$ permettent de reconstruire les fonctions de $\mathcal{K}$. Il est alors naturel d'introduire les *sensing numbers*

$$s_n(\mathcal{K})_V := \inf_{\lambda_1, \ldots, \lambda_n \in V^*} \inf_R \sup_{u \in \mathcal{K}} \|u - R(\lambda_1(u), \ldots, \lambda_n(u))\|_V, \tag{2.126}$$

qui décrivent la meilleure erreur de reconstruction possible uniformément sur $\mathcal{K}$ avec n mesure linéaires optimalement choisies. Un cas important qui rejoint la problématique générale de ce cours est celui où on se restreint à des formes linéaires d'évaluation ponctuelles

$$\lambda_j(u) = u(x^j). \tag{2.127}$$

On obtient alors les *nombres d'échantillonage* (sensing numbers)

$$\rho_n(\mathcal{K})_V := \inf_{x^1,\dots,x^n\in D} \inf_R \sup_{u\in\mathcal{K}} \|u - R(u(x^1),\dots,u(x^n))\|_V. \tag{2.128}$$

On pourra montrer (**exercice**) que si $\mathcal{K}$ est la boule unité de $\mathrm{Lip}([0,1])$ alors $r_n(\mathcal{K})_{L^\infty} \sim n^{-1}$. Indication pour la borne inférieure : montrer que pour tout $x^1,\dots,x^n \in [0,1]$ il existe $u \in \mathcal{K}$ qui s'annulle en tout ces points et telle que $\|u\|_{L^\infty} \geq cn^{-1}$. De même $\rho_n(\mathcal{K})_{L^\infty} \sim n^{-s/d}$ si $\mathcal{K}$ est la boule unité de $\mathcal{C}^s([0,1]^d)$.

3 Reconstruction à partir de données ponctuelles

Nous allons traiter le problème général de la reconstruction d'une fonction u à partir de ses observations exactes ou bruitées en des points $x^1,\dots,x^m \in D$, en cherchant une approximation dans un espace linéaire V_n par des méthodes de moindres carrés. Cette approche nous permettra d'apréhender une notion d'échantillonage optimal, avec des meilleurs résultats que par l'approche de l'interpolation sur laquelle on revient pour commencer.

3.1 Interpolation : constante de Lebesgue et procédés hiérarchiques

Nous avons déja abordé brievement l'interpolation dans un espace $V_n \subset V = \mathcal{C}(D)$ de dimension n, à partir des évaluations en n points $x^1,\dots,x^n$. Le budget d'échantillonage $m = n$ est optimal, mais l'erreur d'interpolation est déteriorée par la constante de Lebesgue suivant

$$\|u - I_n u\|_{L^\infty} \leq (1 + \Lambda_n) e_n(u)_{L^\infty}. \tag{3.1}$$

On a rappelé en particulier que lorsque $V_n = \mathbb{P}_n$ et que l'on travaille sur l'intervalle $D = [-1,1]$ la constante de Lebesgue explose comme 2^n lorsqu'on prend des points uniformes, alors qu'elle augmente lentement comme $\log(n)$ pour les points de Chebychev.

Dans le cas d'espaces V_n plus généraux, la définition de points permettant de limiter croissance de la constante de Lebesgue n'a rien d'évident. Par exemple pour les polynômes à plusieurs variables $V_n = \mathbb{P}_k$ sur avec $n = \binom{k+d}{d}$ sur un domaine $D \subset \mathbb{R}^d$, il n'existe pas d'analogue des points de Chebychev même pour des domaines très simples tels que des polygones en dimension 2.

Il existe cependant une approche systématique permettant d'assurer une croissance au plus linéaire en n. On part d'une base quelconque $\{\phi_1,\dots,\phi_n\}$ de V_n. On note que tout ensemble de points $\{x^1,\dots,x^n\}$ vérifie la propriété d'unisolvence, c'est à dire l'isomorphisme de l'application $L : V_n \to \mathbb{R}^n$

$$v \mapsto Lv := (v(x^1),\dots,v(x^n)), \tag{3.2}$$

si et seulement si la matrice

$$\mathbf{M} = \mathbf{M}_{x^1,\dots,x^n} := (\phi_j(x^i))_{i,j=1,\dots,n}, \tag{3.3}$$

est inversible. En effet si $v = \sum_{j=1}^n a_j \phi_j$ alors $Lv = \mathbf{Ma}$ avec $\mathbf{a} = (a_1, \ldots, a_n)^T$ ce qui nous montre que la propriété d'isomorphisme de $L : V_n \to \mathbb{R}^n$ est équivalente à celle de $\mathbf{M} : \mathbb{R}^n \to \mathbb{R}^n$. La matrice $\mathbf{M}$ est parfois appelée *matrice de collocation*, et c'est celle qui intervient dans le calcul de l'interpolant : si $v = \sum_{j=1}^n a_j \phi_j$ est tel que $v(x^i) = y^i$ pour $i = 1, \ldots, n$ alors les a_j sont solution du système

$$\sum_{j=1}^n a_j \phi_j(x^i) = y^i, \quad i = 1, \ldots, m \quad \iff \quad \mathbf{Ma} = \mathbf{y}, \tag{3.4}$$

où $\mathbf{y} = (y^1, \ldots, y^n)^T$. Pour tout $x \in D$ et $i \in \{1, \ldots, n\}$, introduisons une matrice $\mathbf{M}_i(x)$ définie en remplacant la i-ème ligne $(\phi_1(x^i), \ldots, \phi_n(x^i))$ de $\mathbf{M}$ par $(\phi_1(x), \ldots, \phi_n(x))$, ce qui revient à substituer x à x^i.

Pour tout ensemble unisolvent $\{x^1, \ldots, x^n\}$, on vu que la constante de Lebesgue de l'opérateur d'interpolation est donnée par

$$\Lambda_n = \max_{x \in D} \sum_{i=1}^n |\ell_i(x)|, \tag{3.5}$$

où $\{\ell_1, \ldots, \ell_n\}$ est la base de Lagrange de V_n, définie par les conditions d'interpolation $\ell_i(x^i) = 1$ et $\ell_i(x^j) = 0$ si $j \neq i$. On remarque alors que les ℓ_i peuvent s'exprimer par la formule

$$\ell_i(x) = \frac{\det(\mathbf{M}_i(x))}{\det(\mathbf{M})}. \tag{3.6}$$

En effet, on voit d'une part en développant $\det(\mathbf{M}_i(x))$ suivant la i-ème ligne de la matrice que la fonction $x \mapsto \det(\mathbf{M}_i(x))$ est une combinaison linéaire des ϕ_j et donc appartenant à V_n, et d'autre part on a clairement $\det(\mathbf{M}_i(x^i)) = \det(\mathbf{M})$ et $\det(\mathbf{M}_i(x^j)) = 0$ pour $i \neq j$ (car les lignes i et j sont égales) ce qui montre les conditions d'interpolation.

Bien entendu, la formule (3.6) n'est pas la meilleure façon de calculer les ℓ_i en pratique, mais elle nous indique comment choisir les points $x^1, \ldots, x^n \in D$ pour contrôler la taille des de ces fonctions : on peut les définir afin de maximiser la valeur absolue du determinant de $\mathbf{M}$, c'est à dire comme solution du problème d'optimisation

$$(x1, \ldots, x^n) := \operatorname{argmax}\{|\det(\mathbf{M}_{z^1,\ldots,z^n})| \ : \ (z^1, \ldots, z^n) \in D^n\} \tag{3.7}$$

On pourra vérifier (**exercice**) que les points x^i obtenus ne dépendent pas du choix des fonctions de base ϕ_j qu'on a pris pour définir la matrice de collocation. Un tel choix nous assure que $|\det(\mathbf{M}_i(x))| \leq |\det(\mathbf{M})|$ pour tout i et pour tout $x \in D$, soit

$$\max_{x \in D} |\ell_i(x)| = 1, \tag{3.8}$$

et par conséquent, on a la borne

$$\Lambda_n \leq \sum_{i=1}^n \|\ell_i\|_{L^\infty} = n. \tag{3.9}$$

Il n'existe pas de méthode générale donnant une meilleure constante de Lebesgue pour un espace V_n arbitraire : on peut donc avoir une détérioration multiplicative d'ordre $\mathcal{O}(n)$

entre l'erreur de meilleure approximation et l'erreur d'interpolation (à comparer avec le $\mathcal{O}(\sqrt{n})$ par Kadec-Snobar pour la déterioration de l'erreur par une projection linéaire).

Dans le cas où $V_n = \mathbb{P}_{n-1}$ et $D = [-1, 1]$, en prenant la base des monomes $\phi_j(x) = x^{j-1}$, la matrice de collocation a la forme de Vandermonde

$$\mathbf{M}_{z^1,\dots,z^n} = ((z^i)^{j-1})_{i,j=1,\dots,n}, \tag{3.10}$$

et on a ainsi

$$|\det(\mathbf{M}_{z^1,\dots,z^n})| = \prod_{i\neq j} |z^i - z^j|. \tag{3.11}$$

Dans ce cas les x^i qui maximisent ce produit des écarts mutuels sur tous les n-uplets de $[-1, 1]$ sont appelés *points de Fekete*.

Outre le fait que la constante Λ_n n'a pas une croissance très lente, l'inconvénient de la construction que nous venons de décrire est sa complexité numérique : il faut résoudre un problème d'optimisation à nd variables (les d coordonnées des n points z^i qu'on optimise) et la fonction objectif $\det(\mathbf{M}_{z^1,\dots,z^n})$ n'est ni convexe ni concave, elle possède typiquement un grand nombre de maxima et minima locaux, ce qui met en difficulté tous les algorithmes classiques d'optimisation de type descente de gradient.

Un autre défaut de cette construction est le fait que les points $\{x^1, \dots, x^n\}$ doivent être complètement recalculés si on modifie l'espace V_n. Dans certaines situations, on peut chercher à construire des approximations successives $u_n \in V_n$ pour $n = 0, 1, 2, \dots$ où $(V_n)_{n\geq 0}$ est une suite d'espaces emboités qui pourrait être fixée à l'avance ou construite par un processus adaptatif (par exemple raffinement de maillage, ou ajoût d'une fonction de base bien choisie). Il est alors souhaitable de pouvoir recycler les échantillons $u(x^1), \dots, u(x^n)$ dont on a vu que le calcul pourrait être coûteux, en ajoûtant seulement l'évaluation à un nouveau point x^{n+1} lorsque qu'on passe de V_n à V_{n+1}. Un tel procédé *hierarchique* ou *progressif* signifie donc que les points $\{x^1, \dots, x^n\}$ sont pris la n-ème section d'une unique suite $(x^k)_{k\geq 1}$, ce qui n'est pas le cas dans la construction que nous avons proposé.

Une manière de modifier la construction de façon à ce qu'elle soit hierarchique est de résoudre approximativement le problème d'optimisation par un algorithme dit *greedy* (glouton) : on part de x^1 fixé, et si on a trouvé $\{x^1, \dots, x^{n-1}\}$, on choisit

$$x^n = \text{argmax}\{|\det(\mathbf{M}_{x^1,\dots,x^{n-1},x})| \ : \ x \in D\}. \tag{3.12}$$

Cette approche incrémentale ne conduit bien entendu pas au maximum exact de $|\det(\mathbf{M}_{z^1,\dots,z^n})|$, mais elle possède l'avantage de fournir un procédé hierarchique et une suite de problèmes d'optimisation en dimension d au lieu de nd (mais toujours sujets aux problèmes d'optimum locaux). On peut donner une autre interpretation intuitive de ces points en montrant (**exercice**) que x^n est le point qui maximise l'erreur d'interpolation de ϕ_n aux points $x^1, \dots, x^{n-1}$ dans l'espace V_{n-1} c'est à dire

$$x^n = \text{argmax}_{x\in D}|\phi_n(x) - I_{n-1}\phi_n(x)|. \tag{3.13}$$

Il n'existe malheureusement aucune information sur la constante de Lebesgue Λ_n pour les points $x^1, \dots, x^n$ obtenus par ce procédé hierarchique général, si ce n'est une borne très pessimiste $\Lambda_n \leq 2^{n+1}$ que l'on peut établir par récurrence (**exercice**).

Dans le cas $V_n = \mathbb{P}_{n-1}$ sur $[-1,1]$, la construction revient à choisir x^n maximisant sur $x \in [-1,1]$ le produit $\prod_{j=1}^{n-1} |x - x_j|$, ce qui conduit à la suite des *points de Leja*. On l'initialise typiquement par $x_1 = 1$ et on obtient $x_2 = -1$, $x_3 = 0$, puis x_4 choisi entre $\pm 1/\sqrt{3}$... Pour ces points, on peut vérifier numériquement que la constante de Lebesgue est sous-linéaire $\Lambda_n \leq n$ mais il n'existe à ce jour aucune preuve rigoureuse de ce comportement.

Remarque 3.1 *Toujours pour les polynômes univariés, on a mentionné que la constante de Lebesgue se comporte en $\mathcal{O}(\log n)$ pour les points de Chebychev*

$$x^i = \cos\Big(\frac{2i-1}{2n}\pi\Big), \quad i = 1, \ldots, n \tag{3.14}$$

mais ceux-ci ne sont pas hierarchiques : on obtient des points complètement différents quand on passe de n à $n+1$. Les points de Gauss-Lobatto ou Clemshaw-Curtis qui sont définis par

$$x^i = \cos\Big(\frac{2(i-1)}{2(n-1)}\pi\Big), \quad i = 1, \ldots, n \tag{3.15}$$

sont une variante qui vérifie la propriété d'emboitement lorsqu'on passe de $n = 2^j + 1$ à $n = 2^{j+1} + 1$. Pour ces valeurs particulières de n on a un comportement en $\mathcal{O}(\log n)$ de Λ_n, mais si on cherche à ajoûter des points progressivement de façon à combler les valeurs intermédiaires $2^j + 1 < n < 2^{j+1} + 1$ on constate numériquement que le comportement de Λ_n est beaucoup plus grand pour ces valeurs. Par exemple, lorsqu'on ajoûte les points intermédiaires par ordre croissant sur $[-1,1]$, on constate numériquement une croissance exponentielle de Λ_n en dehors des valeurs $2^j + 1$.

L'ensemble des difficultés que nous venons d'exposer rend les méthodes d'interpolation peu attractives dans des cadres généraux, et tout particulièrement lorsque les données sont bruitées. Les méthodes de moindres carrés vont nous permettre en un certain sens de contourner ces difficultés.

3.2 Méthodes de moindres carrés

On se place dans le cadre du modèle d'observation général

$$y^i = u(x^i) + \eta^i, \quad i = 1, \ldots, m. \tag{3.16}$$

qui est commun au cadre de la régression (les (x^i, y^i) sont distribués dans $D \times \mathbb{R}$ suivant une loi de probabilité qui nous est inconnue) et au cadre de la planification d'expérience ou de l'assimilation de données (nous avons le choix des x^i où on mesure u).

On s'interesse à l'erreur de reconstruction $u - \tilde{u}$ mesurée en norme $L^2(D,\mu)$. Dans le premier cadre μ est la mesure de probabilité marginale de x suivant laquelle les x^i son tirés, et qui nous est inconnue. Dans le deuxième cadre, on peut choisir une mesure μ (pas nécessairement de probabilité) qui nous intéresse pour contrôler l'erreur dans la norme $L^2(D,\mu)$ (par exemple la mesure de Lebesgue). Pour simplifier les notations, on notera systématiquement dans toute la suite de ce chapitre

$$\|v\| = \|v\|_{L^2(D,\mu)} = \Big(\int_D |v(x)|^2 d\mu\Big)^{1/2}, \tag{3.17}$$

et pour un espace V_n de dimension n donné

$$e_n(u) = \min_{v \in V_n} \|u - v\|, \tag{3.18}$$

l'erreur de meilleure approximation dans cette norme.

Pour ce qui est du bruit, $\eta^i = y^i - u(x^i)$ plusieurs modèles peuvent être considérés selon le contexte applicatif :

1. Les η^i sont des réalisations indépendantes d'une variable centrée et de variance κ^2, et sont aussi indépendant des x^i.
2. Les η^i sont variables indépendantes mais leur variance $\kappa^2(x^i)$ dépend du point d'observation suivant une fonction $x \mapsto \kappa^2(x)$. Dans le cas de la régression, la variance moyenne

$$\kappa^2 := \int_D |\kappa(x)|^2 d\mu = \mathbb{E}(|u(x) - y|^2) = \min_v \mathbb{E}(|v(x) - y|^2), \tag{3.19}$$

   représente le risque quadratique minimal atteint par la fonction de régression.
3. Les η^i sont les évaluations d'une fonction de perturbation inconnue : $\eta^i = \eta(x^i)$ où $\eta \in L^2(D, \mu)$.

La meilleure approximation de u dans V_n est donnée par la projection orthogonale

$$P_n u = \operatorname{argmin}\Big\{ \int_D |u(x) - v(x)|^2 d\mu \ : \ v \in V_n \Big\}. \tag{3.20}$$

Son calcul exact est impossible car u n'est mesuré qu'aux point $x^1, \ldots, x^m$ et ces mesures peuvent être bruitées.

On obtient un *estimateur de moindres carrés* en remplacant l'intégrale par une somme discrète :

$$u_n := \operatorname{argmin}\Big\{ \frac{1}{m} \sum_{i=1}^m |y^i - v(x^i)|^2 \ : \ v \in V_n \Big\} \tag{3.21}$$

Dans le cas de la régression, on voit que la quantité qu'on mimimise sur V_n est le *risque empirique* c'est à dire la moyenne empirique du risque $|y - v(x)|^2$.

De façon générale, on parle de *moyenne empirique* d'une variable aléatoire X pour la quantité

$$\tilde{X}_m := \frac{1}{m} \sum_{i=1}^m X_i, \tag{3.22}$$

où les X_i sont des tirages indépendant de X. Une telle quantité est un estimateur sans biais de l'espérance de X puisqu'on a

$$\mathbb{E}(\tilde{X}_m) = m^{-1} \sum_{i=1}^m \mathbb{E}(X_i) = \mathbb{E}(X). \tag{3.23}$$

Si X est de variance $\operatorname{Var}(X) < \infty$, on sait que cet estimateur converge en espérance quadratique au sens où

$$\mathbb{E}(|\tilde{X}_m - \mathbb{E}(X)|^2) = m^{-1} \operatorname{Var}(X) \ \to 0, \tag{3.24}$$

quand $m \to \infty$.

Un estimateur plus général, qui sera particulièrement pertinent dans le cas où on a le choix des points d'évaluations, est obtenu en introduisant des poids dépendant de ces points, c'est à dire en définissant

$$u_n := \operatorname{argmin}\Big\{ \frac{1}{m} \sum_{i=1}^{m} w(x^i)|y^i - v(x^i)|^2 \ : \ v \in V_n \Big\} \tag{3.25}$$

où $w : D \to \mathbb{R}_+$ est une fonction positive que l'on se fixe et qui donne ainsi une importance différente aux échantillons mesurés. On parle de méthode de *moindres carrés à poids*.

Il est facile de montrer qu'il existe toujours une solution au problème de minimisation ci-dessus : on cherche u_n sous la forme $u_n = \sum_{j=1}^{n} a_j L_j$ où $\{L_1, \dots, L_n\}$ est une base de V_n. Le problème revient ainsi à chercher la projection du vecteur des observations $\mathbf{y} = (y^1, \dots, y^m)^T \in \mathbb{R}^m$ sur le sous espace engendré par les vecteurs $\mathbf{f}_j := (L_j(x^1), \dots, L_j(x^m))^T$ pour $j = 1, \dots, n$, pour la norme euclidienne à poids $|z|_w^2 := \sum w(x^j)|z_j|^2$. En écrivant les équations d'orthogonalité, on trouve que le vecteur $\mathbf{a} = (a_1, \dots, a_n)^T$ est solution des *equation normales*

$$\mathbf{G}\mathbf{a} = \mathbf{b}, \tag{3.26}$$

où $\mathbf{b} = (b_1, \dots, b_n)^T$ est le vecteur des produits scalaires

$$b_j = \langle \mathbf{y}, \mathbf{f}_j \rangle_w = \frac{1}{m} \sum_{i=1}^{m} w(x^i) y^i L_j(x^i). \tag{3.27}$$

et $\mathbf{G}$ est la matrice de Gramm

$$\mathbf{G} = (\langle \mathbf{f}_j, \mathbf{f}_k \rangle_w)_{j,k=1,\dots,n}, \quad \langle \mathbf{f}_j, \mathbf{f}_k \rangle_w = \frac{1}{m} \sum_{i=1}^{m} w(x^i) L_j(x^i) L_k(x^i). \tag{3.28}$$

Il faut noter que la solution u_n n'est pas nécessairement unique : par exemple, dans la situation (pathologique mais envisageable) ou toutes les fonctions de V_n s'annullent aux points $x^1, \dots, x^m$ alors n'importe quelle fonction $v \in V_n$ est solution du problème. Bien entendu la solution u_n redevient unique lorsque la matrice $\mathbf{G}$ est inversible. Ceci n'est possible que si $m \geq n$ (**exercice**) hypothèse qu'on fera toujours ici.

Dans le problème des moindres carrés à poids, on a donc remplacé donc la norme $L^2(D, \mu)$ "continue" $\|v\|^2$ par la quantité "discrète"

$$\|v\|_m^2 := \frac{1}{m} \sum_{i=1}^{m} w(x^i)|v(x^i)|^2. \tag{3.29}$$

On parlera de *norme discrète*, même si c'est un abus de language car une fonction v non-nulle pourrait s'annuller en tous les points x^i si bien que $\|v\|_m = 0$. Notons que quand les x^i sont tirés aléatoirement, alors pour une fonction v donnée, $\|v\|_m^2$ est elle même une variable aléatoire. Dans le cas de données non-bruitées

$$y^i = u(x^i), \tag{3.30}$$

on peut donc voir u_n comme la projection orthogonale de u sur V_n au sens de cette norme discrète

$$u_n = P_n^m u := \text{argmin}\{\|u - v\|_m \ : \ v \in V_n\}. \tag{3.31}$$

Cette projection $P_n^m u$ a est elle même un caractère aléatoire (puisqu'elle dépend du tirage des x^i) et on cherchera à comprendre si sa précision que la projection orthogonale $P_n u$ au sens où $\|u - u_n\|$ n'est pas beaucoup plus grand que $e_n(u) = \|u - P_n u\|$, en un sens probabiliste : on peut par exemple chercher à démontre que $\mathbb{E}(\|u - u_n\|) \leq Ce_n(u)$ ou que $\|u - u_n\| \leq Ce_n(x)$ avec grande probabilité. Intuitivement, cela signifie que la norme discrète $\|\cdot\|_m$ doit dans un certain sens être proche de la norme continue $\|\cdot\|$.

Dans le cadre de la régression où les x^i sont tirés aléatoirement et indépendamment suivant la mesure de probabilité μ, il est naturel d'utiliser les moindres carrés sans poids, c'est à dire de prendre $w(x) = 1$, puisqu'alors la norme discrète devient un estimateur sans biais de la norme continue. En effet pour toute fonction v et pour chaque i on a

$$\mathbb{E}(v(x^i)) = \int_D v(x) d\mu, \tag{3.32}$$

ce qui appliqué à $|v|^2$ permet d'obtenir

$$\mathbb{E}(\|v\|_m^2) = \|v\|^2. \tag{3.33}$$

Dans le cadre où on s'est fixé la mesure μ (pas forcément de probabilité) pour quantifier l'erreur suivant la norme $L^2(D, \mu)$ et où on peut choisir les points x^i, on peut faire d'autre choix qui se révèleront plus intéressant du point de vue de la convergence de l'estimateur. En particulier, on peut décider de tirer les x^i indépendamment suivant une mesure de probabilité σ différente de μ. Dans ce cas, il devient intéressant d'introduire une fonction de poid w qui vérifie avec σ la relation

$$w d\sigma = d\mu. \tag{3.34}$$

Lorsque σ et μ sont des mesures avec densités continues $d\sigma = p_\sigma(x)dx$ et $d\mu = p_\mu(x)dx$, cela signifie tout simplement que $wp_\sigma = p_\mu$. Sous une telle condition on aura à nouveau la propriété

$$\mathbb{E}(\|v\|_m^2) = \int_D w(x)|v(x)|^2 d\sigma = \int_D |v(x)|^2 d\mu = \|v\|^2, \tag{3.35}$$

Cette liberté de choix de σ sous la contrainte (3.34), que l'on suppose toujours satisfaite dans la suite de ce chapitre, va nous permettre après une analyse plus poussée d'apréhender la distribution optimale des points x^i qui se traduira par un choix particulier de la mesure d'échantillonage σ. Les résultats qu'on va exposés sont en grande partie associés à la référence de [2] de Cohen-Migliorati citée en introduction.

On sait que pour toute fonction v fixée, les normes $\|v\|_m$ et $\|v\|$ sont "proches" quand $m \to \infty$ au sens où

$$\mathbb{E}(|\|v\|_m^2 - \|v\|^2|^2) = \mathbb{E}\Big(\Big|\frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}(X)\Big|^2\Big) \leq \frac{\text{Var(X)}}{m} \to 0, \tag{3.36}$$

où X est la variable aléatoire donnée par $X = w(x)|v(x)|^2$ lorsque x est distribuée suivant loi σ (en supposant cette variable de variance Var(X) finie), et les X_i sont des tirages indépendants de cette variable.

Afin d'obtenir un résultat d'approximation optimal, il serait idéalement utile que les normes $\|\cdot\|$ et $\|\cdot\|_m$ soient équivalentes, c'est à dire

$$A\|v\| \leq \|v\|_m \leq B\|v\|, \quad v \in L^2(D,\mu), \tag{3.37}$$

pour deux constantes $0 < A \leq B < \infty$. En effet, on aurait alors

$$\|u - u_n\| \leq A^{-1}\|u - u_n\|_m \leq A^{-1}\|u - v\|_m, \quad v \in V_n, \tag{3.38}$$

et donc en particulier

$$\|u - u_n\| \leq A^{-1}\|u - P_n u\|_m \leq BA^{-1}\|u - P_n u\| = Ce_n(u), \tag{3.39}$$

avec $C = B/A$. Il est cependant immédiat de voir que l'équivalence (3.37) ne peut être vraie, même lorsque m devient très grand : on peut toujours trouver une fonction $v \in L^2(D,\mu)$ non-nulle mais qui s'annulle en tous les points $x^1, \ldots, x^m$, ce qui contredit (3.37).

On va se contenter de rechercher une propriété plus faible : l'équivalence des deux normes lorsqu'on se restreint à l'espace V_n. On cherchera cette équivalence sous la forme

$$(1-\delta)\|v\|^2 \leq \|v\|_m^2 \leq (1+\delta)\|v\|^2, \quad v \in V_n, \tag{3.40}$$

avec $\delta \in]0,1[$. Pour comprendre pourquoi une telle propriété peut être réalisée, considérons l'exemple très simple où V_n est l'espace des fonctions constantes par morceaux sur une partition $\{D_1, \ldots, D_n\}$ de D. Dans ce cas on voit que si $v \in V_n$ a pour valeur v_i sur D_i, on a

$$\|v\|^2 = \sum_{i=1}^{n} \mu(D_i)|v_i|^2, \tag{3.41}$$

et par conséquent si on prend $m = n$ points $x^1, \ldots, x^n$ tels que $x^i \in D_i$, et $w(x^i) = m\mu(D_i)$ on a alors exactement $\|v\|^2 = \|v\|_m^2$, c'est à dire l'équivalence (3.40) avec $\delta = 0$. Notons que la propriété (3.40) est un événement E_δ qui dépend du tirage indépendant des $x^1, \ldots, x^m$ suivant la loi σ, et se produit donc avec une certaine probabilité $\Pr(E_\delta)$ que nous allons étudier.

L'équivalence (3.40) a une interprètation en termes matriciels qui se révelera particulièrement utile. Supposons que $\{L_1, \ldots, L_n\}$ soit une base de V_n qui est orthonormée pour le produit scalaire $L^2(D,\mu)$. Alors pour tout $v = \sum_{j=1}^n a_j L_j$, on a

$$\|v\|^2 = \sum_{j=1}^{n} |a_j|^2 = |\mathbf{a}|^2, \quad \mathbf{a} = (a_1, \ldots, a_n)^T. \tag{3.42}$$

et d'autre part

$$\|v\|_m^2 = \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \langle L_j, L_k\rangle_m = \sum_{j=1}^{n}\sum_{k=1}^{n} G_{j,k} a_j a_k = \mathbf{a}^T \mathbf{G}\mathbf{a}, \tag{3.43}$$

où $\mathbf{G}$ est la matrice de Gramm d'éléments $G_{j,k} = \langle L_j, L_k \rangle_m = \frac{1}{m} \sum_{i=1}^m w(x^i) L_j(x^i) L_k(x^i)$. Par conséquent, la propriété (3.40) peut aussi s'écrire

$$(1-\delta)|\mathbf{a}|^2 \leq \mathbf{a}^T \mathbf{G} \mathbf{a} \leq (1+\delta)|\mathbf{a}|^2, \quad \mathbf{a} \in \mathbb{R}^n, \tag{3.44}$$

ce qui signifie pour les valeurs propres de $\mathbf{G}$,

$$\lambda_{\min}(\mathbf{G}) \geq 1-\delta \quad \text{et} \quad \lambda_{\max}(\mathbf{G}) \leq 1+\delta, \tag{3.45}$$

autrement dit $(1-\delta)\mathbf{I} \leq \mathbf{G} \leq (1+\delta)\mathbf{I}$ au sens des matrices symmétriques. Notons que cette propriété implique en particulier que le nombre de conditionnement de la matrice $\mathbf{G}$ est controlé par $\frac{1+\delta}{1-\delta}$ ce qui est important pour la stabilité du calcul numérique de l'estimateur par le système des équations normales (3.26). Notons que (3.45) équivaut aussi à la propriété

$$\|\mathbf{G} - \mathbf{I}\|_2 \leq \delta, \tag{3.46}$$

où $\|\mathbf{M}\|_2 := \max\{|\mathbf{M}\mathbf{a}| \, : \, |\mathbf{a}| = 1\}$ désigne la norme spectrale des matrices $n \times n$. On note que $\mathbb{E}(G_{j,k}) = \langle L_j, L_k \rangle = \delta_{j,k}$ c'est à dire $\mathbb{E}(\mathbf{G}) = \mathbf{I}$. L'étude de l'équivalence de norme uniforme sur V_n revient donc à étudier la "concentration" de la variable aléatoire matricielle $\mathbf{G}$ autour de son espérance en norme spectrale.

Montrons par un premier calcul simple en quel sens l'équivalence de norme (3.40) peut nous conduire à une borne d'erreur optimale pour l'estimateur des moindres carrés, dans le cas de données non bruitées. Dans ce cas $u_n = P_n^m u$ est la projection orthogonale de u sur V_n pour la norme $\|\cdot\|_m$. On écrit d'abord que sous l'évenement E_δ,

$$\|u - u_n\|^2 = \|u - P_n u\|^2 + \|P_n u - u_n\|^2 \leq \|u - P_n u\|^2 + (1-\delta)^{-1} \|P_n u - u_n\|_m^2, \tag{3.47}$$

où on a utilisé Pythagore et (3.40) puisque $P_n u - u_n \in V_n$. Puis toujours par Pythagore,

$$\|P_n u - u_n\|_m^2 = \|P_n u - P_n^m u\|_m^2 = \|u - P_n u\|_m^2 - \|u - P_n^m u\|_m^2 \leq \|u - P_n u\|_m^2. \tag{3.48}$$

On a ainsi établi

$$\|u - u_n\|^2 \leq \|u - P_n u\|^2 + (1-\delta)^{-1} \|u - P_n u\|_m^2 = e_n(u)^2 + (1-\delta)^{-1} \|u - P_n u\|_m^2. \tag{3.49}$$

En notant χ_{E_δ} l'indicatrice de l'évenement E_δ, on a donc

$$\mathbb{E}(\|u-u_n\|^2 \chi_{E_\delta}) \leq e_n(u)^2 + (1-\delta)^{-1} \mathbb{E}(\|u - P_n u\|_m^2) = e_n(u)^2 + (1-\delta)^{-1} \|u - P_n u\|^2 = C e_n(u)^2, \tag{3.50}$$

avec $C = 1 + (1-\delta)^{-1}$.

Notons qu'il est facile de tester numériquement si l'évenement E_δ est vérifié : il suffit de calculer les valeurs propres de $\mathbf{G}$ qui doivent être comprises entre $1-\delta$ et $1+\delta$. Si ce n'est pas le cas, on peut décider que l'estimateur des moindres carrés n'est pas fiable. On défini ainsi un estimateur conditionné à E_δ par

$$\tilde{u} = u_n \chi_{E_\delta}, \tag{3.51}$$

c'est à dire en prenant simplement $\tilde{u} = 0$ lorsque E_δ n'est pas vérifié. On a alors

$$\mathbb{E}(\|u - \tilde{u}\|^2) = \mathbb{E}(\|u - u_n\|^2 \chi_{E_\delta}) + \mathbb{E}(\|u\|^2 \chi_{E_\delta^c}) \leq C e_n(u)^2 + \Pr(E_\delta^c) \|u\|^2. \tag{3.52}$$

On voit ainsi que l'erreur d'approximation est presque optimale *en espérance quadratique moyenne* avec une constante multiplicative $C = 1 + (1-\delta)^{-1}$ et à l'addition près d'un terme lié à la probabilité que l'évenement E_δ n'ait pas lieu. Notre étude qui va suivre va donc consister à comprendre si on peut rendre cette probabilité très petite en prenant m suffisament grand mais si possible en ne s'éloignant pas trop du budget optimal $m = n$. Notons qu'il n'est pas particulièrement utile d'exiger une valeur de δ très faible pour avoir une constante multiplicative raisonnable. Le choix $\delta = \frac{1}{2}$ conduit à $C = 3$. Nous verrons par une analyse plus fine qu'il est en fait possible d'obtenir une constante multiplicative C arbitrairement proche de 1.

3.3 Fonction de Christoffel inverse et inégalités de concentration

L'étude de l'équivalence entre les normes $\|\cdot\|$ et $\|\cdot\|_m$ sur V_n va faire appel à deux outils fondamentaux que nous développons dans cette section.

Le premier est une grandeur qui permet de relier le comportement ponctuel d'une fonction de V_n et sa norme $L^2(D,\mu)$. Si $\{L_1,\dots,L_n\}$ est une base orthonormée de V_n pour cette norme, on introduit la fonction

$$x \mapsto k_n(x) := \sum_{j=1}^{n} |L_j(x)|^2. \tag{3.53}$$

Notons que cette fonction est en fait indépendante du choix de la base orthonormée de V_n puisque la quantité $k_n(x)$ reste inchangée si on applique une matrice unitaire au vecteur $(L_1(x),\dots,L_n(x))^T$. Elle ne dépend que de V_n et de la mesure μ.

On remarque que si $v = \sum_{j=1}^n a_j L_j$ est une fonction quelconque de V_n on a pour tout point $x \in D$,

$$|v(x)|^2 = \Big|\sum_{j=1}^{n} a_j L_j(x)\Big|^2 \le \Big(\sum_{j=1}^{n} |a_j|^2\Big)\Big(\sum_{j=1}^{n} |L_j(x)|^2\Big) = k_n(x)\|v\|^2, \tag{3.54}$$

D'autre part, pour tout $x \in D$ fixé il existe toujours une fonction $v \in V_n$ telle que l'inégalité de Cauchy-Schwarz utilisée soit une égalité, ce qui nous montre $k_n(x)$ est la plus petite constante telle que l'inégalité ci-dessus est vérifiée

$$k_n(x) = \max_{v\in V_n} \frac{|v(x)|^2}{\|v\|^2} = \max\{|v(x)|^2 \,:\, v \in V_n,\, \|v\| = 1\}. \tag{3.55}$$

Dans la littérature mathématique, on considère souvent la grandeur inverse

$$x \mapsto \lambda_n(x) := k_n^{-1}(x) = \min_{v\in V_n} \frac{\|v\|^2}{|v(x)|^2} = \min\{\|v\|^2 \,:\, v \in V_n,\, v(x) = 1\}, \tag{3.56}$$

qui est appellée *fonction de Christoffel* pour l'espace V_n et la mesure μ. Pour cette raison, la fonction k_n est parfois appellée fonction de Christoffel inverse.

Lorsque les fonctions des V_n appartiennent à $L^\infty(D)$, nous utiliserons aussi la notation

$$K_n = \|k_n\|_{L^\infty} = \sup_{x\in D} \sum_{j=1}^{n} |L_j(x)|^2, \tag{3.57}$$

grandeur qui peut donc être aussi définie par

$$K_n = \max_{v \in V_n} \frac{\|v\|_{L^\infty}^2}{\|v\|^2}, \tag{3.58}$$

c'est à dire la plus petite constante K telle que $\|v\|_{L^\infty}^2 \leq K\|v\|^2$ pour tout $v \in V_n$. Notons que si μ est une mesure de masse finie, l'inégalité dans l'autre sens $\|v\|^2 \leq \mu(D)\|v\|_{L^\infty}^2$ est vérifiée pour toute fonction $v \in L^\infty(D)$.

Notons que par construction

$$\int_D k_n(x) d\mu = \sum_{j=1}^{n} \|L_j\|^2 = n. \tag{3.59}$$

Par conséquent si μ est une mesure de probabilité on a toujours

$$K_n = \|k_n\|_{L^\infty} \geq \int_D k_n(x) d\mu = n. \tag{3.60}$$

Donnons quelques exemple simples pour lesquels on peut précisément estimer K_n :

1. Polynômes trigonométriques : on prend $D = [0, 1]$ avec $d\mu = dx$ la mesure de Lebesgue, et l'espace V_n engendré par les fonctions $x \mapsto e_j(x) = \exp(i2\pi jx)$ pour $j = -k, \ldots, k$ avec $n = 2k+1$ qui forment une base orthonormale. Comme $|e_j|^2 = 1$, on trouve dans ce cas que k_n est constante, égale à $2k + 1 = n$, et en particulier la valeure minimale $K_n = n$ est atteinte.

2. Fonctions constantes par morceaux : on considère un domaine D muni d'une mesure de probablité μ et V_n l'espace des fonctions constantes par morceaux sur une partition $\{D_1, \ldots, D_n\}$, dont une base orthonormée est donnée par les fonctions $L_j = \mu(D_j)^{-1/2}\chi_{D_j}$. Ainsi on a

$$K_n = \max \mu(D_j)^{-1}. \tag{3.61}$$

   On voit en particulier que $K_n = n$ si on a choisit une partition μ-uniforme, au sens où $\mu(D_j) = 1/n$ pour $j = 1, \ldots, n$. Mais pour des partitions générales on aura $K_n > n$.

3. Polynômes de Legendre : on considère $D = [-1, 1]$ muni de la mesure de probabilité uniforme $d\mu = \frac{1}{2}dx$ et $V_n = \mathbb{P}_{n-1}$ l'espace des polynômes algébriques de degré $n - 1$. Une base orthogonale est donnée par les polynômes de Legendre L_j pour $j = 0, \ldots, n - 1$. Ceux-ci sont habituellement normalisés en norme max $\|L_j\|_{L^\infty} = L_j(1) = 1$, et après renormalisation dans $L^2(D, d\mu)$ leur valeur maximale est toujours au point 1 mais égale à $(2j + 1)^{1/2}$. On a ainsi

$$K_n = \sum_{j=0}^{n-1} (1 + 2j) = n + 2n(n - 1)/2 = n^2, \tag{3.62}$$

   ce qui nous montre que dans ce cas K_n augmente plus rapidement que n.

4. Polynômes de Chebychev : on considère $D = [-1, 1]$ et $V_n = \mathbb{P}_{n-1}$ l'espace des polynômes algébriques de degré $n-1$. Les polynômes de Chebychev définis par $T_j(\cos(t)) = \cos(jt)$ forment une base de V_n orthogonale dans $L^2(D, \mu)$ pour la mesure de probabilité $d\mu = \frac{dx}{\pi\sqrt{1-x^2}}$. On la normalise dans $L^2(D, \mu)$ en multipliant T_j par $\sqrt{2}$ pour $j \geq 1$ et en conservant T_0, et on obtient ainsi

$$K_n = 1 + \sum_{j=1}^{n-1} 2 = 2n - 1, \tag{3.63}$$

ce qui nous montre que dans ce cas K_n se comporte en $\mathcal{O}(n)$.

Pour l'analyse des méthodes des moindres carrés à poids, il sera utile de considérer les quantités plus générales définie par

$$k_{n,w}(x) := w(x)k_n(x), \tag{3.64}$$

et

$$K_{n,w} = \|k_{n,w}\|_{L^\infty} := \sup_{x \in D} w(x) \sum_{j=1}^{n} |L_j(x)|^2. \tag{3.65}$$

Rappelons que la fonction de poids w est choisi de manière à ce que $wd\sigma = d\mu$ où σ est la mesure de probabilité avec laquelle on échantillone les x^i. Ceci nous montre que l'on a toujours

$$\int_D k_{n,w}(x)d\sigma = \int_D k_n(x)d\mu = n, \tag{3.66}$$

et par conséquent la borne inférieure

$$K_{n,w} \geq \int_D k_{n,w}(x)d\sigma = n. \tag{3.67}$$

Notons qu'un choix particulier de σ et w permet d'atteindre cette borne inférieure : il suffit de considérer la mesure de probabilité définie par

$$d\sigma := \frac{k_n}{n}d\mu, \tag{3.68}$$

qui est bien de masse $\int_D d\sigma = 1$ puisque $\int_D k_n d\mu = n$, et de définir le poids

$$w = \frac{n}{k_n}, \tag{3.69}$$

afin de vérifier la relation de compatibilité $wd\sigma = d\mu$. On voit que pour ce choix on a toujours $k_{n,w} = n$ et $K_{n,w} = n$. Nous verrons qu'il s'agit d'un choix optimal au sens où il permet d'établir la validité de l'équivalence (3.40) avec une grande probabilité, et avec un budget d'échantillonage m qui ne sera pas beaucoup plus élevé que n.

Le deuxième outil central à notre étude sera un ensemble de techniques pour quantifier de manière probabilisite la concentration d'une moyenne empirique autour de son

espérance. Rappelons que pour toute fonction v, on peut introduire la variable aléatoire $X = w(x)|v(x)|^2$ où x est une variable distribuée suivant la loi σ, de sorte que

$$\|v\|_m^2 = \frac{1}{m}\sum_{i=1}^m X_i \quad \text{et} \quad \|v\|^2 = \mathbb{E}(X). \tag{3.70}$$

De façon générale, si $X_1, \ldots, X_m$ sont m tirages indépendant d'une variable aléatoire X réelle on cherche pour $\delta > 0$ à estimer par au dessus la probabilité

$$P(\delta) = \Pr\Big(\Big|\frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}(X)\Big| \geq \delta\Big). \tag{3.71}$$

Une telle estimation s'appelle *inégalité de concentration*. Si X est de variance $\tau^2 = \mathrm{Var}(X)$ finie, alors on sait que

$$\mathbb{E}\Big(\Big|\frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}(X)\Big|^2\Big) = \frac{\tau^2}{m}. \tag{3.72}$$

On peut ainsi utiliser l'inégalité de Markov qui nous dit que si Y est une variable aléatoire positive et $\delta > 0$, alors $\Pr(Y > \delta) \leq \delta^{-1}\mathbb{E}(Y)$. Ceci entraîne en particulier $\Pr(|Z| > \delta) \leq \delta^{-2}\mathbb{E}(|Z|^2)$ pour toute variable aléatoire Z de variance finie. Ainsi en posant

$$Z = \frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}(X), \tag{3.73}$$

on trouve

$$P(\delta) = \Pr\{|Z| \geq \delta\} \leq \frac{\tau^2}{m\delta^2} \tag{3.74}$$

Cette première estimation élémentaire se révèle souvent insuffisante et peut être grandement améliorée en tirant parti du phénomène de "concentration gaussienne" que suggère la loi forte des grands nombre : on sait que la variable renormalisée $\tau^{-1}\sqrt{m}Z$ converge en loi vers une gaussienne normale centrée $\mathcal{N}(0,1)$, c'est à dire de distribution

$$g(s) = \frac{1}{\sqrt{2\pi}}\exp(-s^2/2). \tag{3.75}$$

Si $\tau^{-1}\sqrt{m}Z$ suivait exactement cette loi gaussienne (cela serait le cas si X était elle même une variable gaussienne) alors on pourrait écrire

$$P(\delta) = \Pr(|Z| \geq \delta) = \Pr(|\tau^{-1}\sqrt{m}Z| \geq \tau^{-1}\sqrt{m}\delta) = \frac{1}{\sqrt{2\pi}}\int_{|s|>\tau^{-1}\sqrt{m}\delta} \exp(-s^2/2)ds. \tag{3.76}$$

Le reste de la gaussienne décroit lui même exponentiellement : par exemple si $S \geq 1$, on peut simplement écrire

$$\int_{s>S} \exp(-s^2/2)ds \leq \int_{s>S} s\exp(-s^2/2)ds = \exp(-S^2/2), \tag{3.77}$$

ce qui nous montre que pour $\delta \geq \tau/\sqrt{m}$ on aurait ainsi

$$P(\delta) \leq \sqrt{2/\pi}\exp\Big(-\frac{m\delta^2}{\tau^2}\Big), \tag{3.78}$$

qui serait une bien meilleure estimation que (3.74). Nous allons montrer qu'une telle estimation peut-être obtenue pour une classe de variables aléatoires qui généralise les variables gaussiennes.

Définition 3.1 *Une variable Y centrée est dite sous-gaussienne de facteur de variance $v > 0$ si et seulement si*

$$\mathbb{E}(\exp(tY)) \leq \exp\Big(\frac{vt^2}{2}\Big), \tag{3.79}$$

pour tout $t \in \mathbb{R}$.

Cette définition est justifiée par le cas particulier des variables gaussiennes : si Y est une gaussienne centrée de variance τ^2, on a

$$\begin{aligned}\mathbb{E}(\exp(tY)) &= \frac{1}{\tau\sqrt{2\pi}}\int_{\mathbb{R}}\exp(ty)\exp\Big(-\frac{y^2}{2\tau^2}\Big)dy\\ &= \frac{1}{\tau\sqrt{2\pi}}\int_{\mathbb{R}}\exp\Big(-\frac{y^2-2\tau^2ty}{2\tau^2}\Big)dy\\ &= \exp\Big(\frac{\tau^2t^2}{2}\Big)\frac{1}{\tau\sqrt{2\pi}}\int_{\mathbb{R}}\exp\Big(-\frac{(y-\tau^2t)^2}{2\tau^2}\Big)dy = \exp\Big(\frac{\tau^2t^2}{2}\Big).\end{aligned}$$

Dans ce cas le facteur de variance est donc simplement $v = \tau^2$.

Un autre cas particulièrement important de variables sous-gaussienne est celui des variables bornées : si Y est une variable centrée telle que

$$|Y| \leq K, \tag{3.80}$$

presque sûrement, on peut écrire

$$\mathbb{E}(\exp(tY)) = \int_{-K}^{K}\exp(tx)p(x)dx, \tag{3.81}$$

où p est la densité de probabilité de Y. En posant $x = sK + (s-1)K$ avec $s \in [0,1]$ et $q(s) = p(x)$, il vient

$$\mathbb{E}(Y\exp(tY)) = 2K\int_0^1\exp(stK+(s-1)tK)q(s)ds \leq 2K\int_0^1[s\exp(tK)+(1-s)\exp(-tK)]q(s)ds. \tag{3.82}$$

Les égalités $\mathbb{E}(Y) = \int_{-K}^{K}xp(x)dx = 0$ et $\int_{-K}^{K}p(x)dx = 1$ entraînent $\int_0^1 sq(s)ds = \int_0^1(1-s)q(s)ds = \frac{1}{4K}$, et par conséquent

$$\mathbb{E}(\exp(tY)) \leq \frac{1}{2}(\exp(tK)+\exp(-tK)) \leq \exp\Big(\frac{t^2K^2}{2}\Big), \tag{3.83}$$

la deuxième inégalité pouvant se vérifier par comparaison des développement limités. Les variables centrées bornées par K sont donc sous-gaussiennes de facteur de variance $K^2/2$.

Revenons à l'estimation de $P(\delta) = \Pr\{|Z| \geq \delta\}$, où $Z = \frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}(X)$. En posant $Y = X - \mathbb{E}(X)$ on se ramène à une variable centrée, avec $Z = \frac{1}{m}\sum_{i=1}^m Y_i$. Supposons que Y soit sous-gaussienne de facteur de variance v. On va évaluer séparément les deux termes

$$P(\delta) = P(Z > \delta) + P(-Z > \delta), \tag{3.84}$$

en utilisant la stratégie suivante due à Chernoff : pour tout $t \geq 0$, on peut écrire

$$\Pr(Z > \delta) = \Pr\Big(\sum_{i=1}^m tY_i \geq tm\delta\Big) = \Pr\Big(\exp\Big(\sum_{i=1}^m tY_i\Big) \geq \exp(tm\delta)\Big), \tag{3.85}$$

où on a utilisé le caractère croissant de l'exponentielle. On utilise alors l'égalité de Markov qui nous donne

$$\Pr(Z > \delta) \leq \exp(-tm\delta)\,\mathbb{E}\Big(\exp\Big(\sum_{i=1}^m tY_i\Big)\Big) = \exp(-tm\delta)\,\mathbb{E}\Big(\prod_{i=1}^m \exp(tY_i)\Big). \tag{3.86}$$

Par l'indépendance des Y_i et le caractère sous-gaussien, on obtient

$$\Pr(Z > t) \leq \Big(\exp(-t\delta)\,\mathbb{E}(\exp(tY))\Big)^m \leq \exp\Big(m\Big(\frac{1}{2}vt^2 - t\delta\Big)\Big) \tag{3.87}$$

On choisit finalement la valeur de t minimisant $\frac{1}{2}vt^2 - t\delta$ soit $t = \frac{\delta}{v}$, ce qui donne la borne

$$\Pr(Z > \delta) \leq \exp\Big(-\frac{m\delta^2}{2v}\Big). \tag{3.88}$$

La probabilité $\Pr(Z < -\delta) = \Pr(-Z > \delta)$ vérifie la même borne (il suffit de considérer la variable $-Y$), et on obtient ainsi l'inégalité de concentration suivante.

Théorème 3.1 *Si $Y = X - \mathbb{E}(X)$ est une variable sous-gaussienne de facteur de variance v, on a*

$$P(\delta) = \Pr\Big(\Big|\frac{1}{m}\sum_{i=1}^m X_i - \mathbb{E}(X)\Big| \geq \delta\Big) \leq 2\exp\Big(-\frac{m\delta^2}{2v}\Big). \tag{3.89}$$

En particulier si Y est bornée suivant $|Y| \leq K$, on a l'inégalité de Hoeffding

$$P(\delta) \ \leq 2\exp\Big(-\frac{m\delta^2}{K^2}\Big). \tag{3.90}$$

Revenons à notre problème de départ qui est d'étudier la probabilité de l'évènement E_δ, c'est à dire l'équivalence (3.40) entre $\|v\|_m^2$ et $\|v\|^2$ uniformément sur $v \in V_n$. On note que cette équivalence peut aussi s'écrire sous la forme

$$\Big|\frac{1}{\|v\|^2}\frac{1}{m}\sum_{i=1}^m w(x^i)|v(x^i)|^2 - 1\Big| \leq \delta. \tag{3.91}$$

Etudions tout d'abord sa validité *pour une seule fonction* $v \in V_n$ que l'on fixe. On peut dans ce cas introduire la variable

$$X = \frac{w(x)|v(x)|^2}{\|v\|^2}, \tag{3.92}$$

où x est une variable distribuée suivant la loi σ. On a ainsi $\mathbb{E}(X) = 1$ et on peut écrire

$$\Pr\Big(\Big|\frac{1}{\|v\|^2}\sum_{i=1}^{m} w(x^i)|v(x^i)|^2 - 1\Big| \leq \delta\Big) = 1 - P(\delta), \tag{3.93}$$

où $P(\delta) = \Pr(|Z| > \delta)$. Notons que puisque X est une variable positive, alors si $0 \leq X \leq K$ la variable recentrée $Y = X - \mathbb{E}(X)$ vérifie

$$|Y| \leq K. \tag{3.94}$$

D'autre part, on sait que pour tout $x \in D$ on a $\frac{|v(x)|^2}{\|v\|^2} \leq k_n(x)$, et par conséquent $X \leq k_{n,w}(x)$, ce qui nous montre que Y vérifie la borne

$$|Y| \leq K, \quad K = K_{n,w}. \tag{3.95}$$

L'inégalité de Hoeffding nous donne par conséquent

$$P(\delta) \leq 2\exp\Big(-\frac{m\delta^2}{K_{n,w}^2}\Big). \tag{3.96}$$

Cette première estimation présente l'intérêt de nous montrer le rôle de la quantité $K_{n,w}$ pour comprendre le budget d'échantillonage nécessaire pour rendre $P(\delta)$ petit. On a vu qu'on peut prendre par exemple $\delta = \frac{1}{2}$, et on obtient alors

$$P(1/2) \leq 2\exp\Big(-\frac{m}{4K_{n,w}^2}\Big). \tag{3.97}$$

Ceci nous montre que pour tout $\varepsilon > 0$, on a $P(1/2) < \varepsilon$ dès qu'on est dans le régime d'échantillonage

$$m \geq 4K_{n,w}^2 \ln(2/\varepsilon). \tag{3.98}$$

Cette première estimation n'est pas très bonne car on voit d'une part qu'au facteur logarithmique près, elle va imposer un budget d'échantillonage

$$m \geq K_{n,w}^2 \geq n^2, \tag{3.99}$$

qui est donc largement sous optimal. Et d'autre part, on a simplement montré la grande probabilité de l'évenement

$$E_{\delta,v} = \{(1-\delta)\|v\|^2 \leq \|v\|_m^2 \leq (1+\delta)\|v\|^2\}, \tag{3.100}$$

pour une seule fonction $v \in V_n$ fixée, pour lequel on a établi

$$P(\delta) = \Pr(E_{\delta,v}^c) \leq 2\exp\Big(-\frac{m\delta^2}{K_{n,w}^2}\Big). \tag{3.101}$$

Il nous faudra des outils supplémentaires pour établir la grande probabilité de l'évènement uniforme

$$E_\delta = \{(1-\delta)\|v\|^2 \leq \|v\|_m^2 \leq (1+\delta)\|v\|^2 \ : \ v \in V_n\} = \bigcap_{v\in V_n} E_{\delta,v}, \tag{3.102}$$

autrement dit majorer la probabilité de l'évenement complémentaire

$$E_\delta^c = \{\exists v \in V_n \ : \ |\,\|v\|^2 - \|v\|_m^2| > \delta\|v\|^2\} = \bigcup_{v\in V_n} E_{\delta,v}^c, \tag{3.103}$$

Nous allons dans un premier temps montrer que le régime $m \geq K_{n,w}^2$ peut être remplacé par $m \geq K_{n,w}$, en utilisant une autre inégalité de concentration. Supposons que X est une variable positive bornée par $0 \leq X \leq K$, et telle que $\mathbb{E}(X) = 1$. Pour $0 < \delta < 1$, nous allons estimer

$$\begin{aligned} P(\delta) \ &= \Pr\Big(|\frac{1}{m}\sum_{i=1}^m X_i - 1| > \delta\Big) \\ &= \Pr\Big(\frac{1}{m}\sum_{i=1}^m X_i > 1+\delta\Big) + \Pr\Big(\frac{1}{m}\sum_{i=1}^m X_i < 1-\delta\Big), \end{aligned}$$

en traitant les deux termes de manières un peu différente de celle utilisée précèdemment. Pour le premier terme, on écrit pour tout $t > 0$,

$$\begin{aligned} \Pr\Big(\frac{1}{m}\sum_{i=1}^m X_i > 1+\delta\Big) \ &= \Pr\Big(\exp\Big(\sum_{i=1}^m tX_i\Big) > \exp(mt(1+\delta))\Big) \\ &\leq \Big(\exp(-t(1+\delta))\,\mathbb{E}(\exp(tX))\Big)^m, \end{aligned}$$

où on a utilisé l'inégalité de Markov comme précédemment. La convexité de l'exponentielle nous donne

$$\exp(tX) \leq 1 + \frac{X}{K}(\exp(tK) - 1) \tag{3.104}$$

et par conséquent

$$\mathbb{E}(\exp(tX)) \leq 1 + K^{-1}(\exp(tK) - 1). \tag{3.105}$$

En prenant la valeur particulière $t = K^{-1}\ln(1+\delta) > 0$, on trouve

$$\mathbb{E}(\exp(tX)) \leq 1 + K^{-1}\delta \leq \exp(K^{-1}\delta), \tag{3.106}$$

ce qui nous conduit finalement à

$$\Pr\Big(\frac{1}{m}\sum_{i=1}^m X_i > 1+\delta\Big) \leq \exp\Big(-\frac{mc_\delta}{K}\Big), \quad c_\delta := (1+\delta)\ln(1+\delta) - \delta > 0. \tag{3.107}$$

Pour estimer le deuxième terme, on procède de façon similaire en écrivant

$$\Pr\Big(\frac{1}{m}\sum_{i=1}^m X_i < 1-\delta\Big) \leq \Big(\exp(t(1-\delta))\,\mathbb{E}(\exp(-tX))\Big)^m, \tag{3.108}$$

puis $\mathbb{E}(e^{-tX}) \leq 1 + K^{-1}(e^{-tK} - 1)$ et en prenant $t = -K^{-1}\ln(1-\delta)$, ce qui conduit à

$$\Pr\Big\{\frac{1}{m}\sum_{i=1}^{m} X_i < 1-\delta\Big\} \leq \exp\Big(-\frac{m\bar{c}_\delta}{K}\Big), \quad \bar{c}_\delta := \delta + (1-\delta)\ln(1-\delta) > 0. \tag{3.109}$$

On peut vérifier que $\bar{c}_\delta \geq c_\delta$ et ceci nous conduit ainsi à l'inégalité de Chernoff

$$P(\delta) = \Pr\Big(\Big|\frac{1}{m}\sum_{i=1}^{m} X_i - 1\Big| > \delta\Big) \leq 2\exp\Big(-\frac{mc_\delta}{K}\Big). \tag{3.110}$$

Appliquée à la variable $X = \frac{w(x)|v(x)|^2}{\|v\|^2}$ pour $v \in V_n$, cette inégalité nous donne ainsi

$$P(\delta) = \Pr(E_{\delta,v}^c) \leq 2\exp\Big(-\frac{mc_\delta}{K_{n,w}}\Big). \tag{3.111}$$

Ceci nous montre que pour tout $\varepsilon > 0$, on a $P(\delta) < \varepsilon$ dès qu'on est dans le régime d'échantillonage

$$m \geq \frac{1}{c_\delta} K_{n,w}\ln(2/\varepsilon), \tag{3.112}$$

qui est plus favorable que celui qu'on a ait obtenu précèdemment puisque $K_{n,w}$ n'est plus élevé au carré. Pour $\delta = \frac{1}{2}$ on trouve $\frac{1}{c_{1/2}} \approx 10$.

3.4 Inégalités de concentration uniformes et matricelles

On cherche maintenant à borner la probabilité de l'évènement E_δ^c, c'est à dire établir avec grande probabilité la validité de l'équivalence des normes $\|\cdot\|$ et $\|\cdot\|_m$ uniformément sur toutes les fonctions $v \in V_n$. Notons que si l'on considère un ensemble fini de fonctions $S = \{v_1, \ldots, v_k\} \subset V_n$ et l'évenement

$$E_{\delta,S} := \{(1-\delta)\|v\|^2 \leq \|v\|_m^2 \leq (1+\delta)\|v\|^2 \ : \ v \in S\} = \bigcap_{v\in S} E_{\delta,v}, \tag{3.113}$$

on peut estimer la probabilité de $E_{\delta,S}^c = \bigcup_{v\in S} E_{\delta,v}^c$ par la borne d'union

$$\Pr(E_{\delta,S}^c) \leq \sum_{v\in S} \Pr(E_{\delta,v}^c) \leq 2\#(S)\exp\Big(-\frac{mc_\delta}{K_{n,w}}\Big). \tag{3.114}$$

On peut ainsi controler la probabilité de l'équivalence des normes uniformément sur un ensemble fini de fonctions, mais on ne peut pas pousser ce raisonnement jusqu'à V_n tout entier puisque c'est un ensemble de cardinal infini.

Afin d'établir une estimation de $\Pr(E_\delta^c)$, une première idée consiste à utiliser une technique qui permet de se ramener à un ensemble fini de fonctions. On remarque tout d'abord que l'équivalence sur V_n peut s'exprimer à travers l'opérateur $T : \mathbb{R}^n \to V_n$ défini

$$T\mathbf{a} = \sum_{j=1}^{n} a_j L_j, \quad \mathbf{a} = (a_1, \ldots, a_n)^T, \tag{3.115}$$

en écrivant

$$(1-\delta)|\mathbf{a}|^2 \leq \|T\mathbf{a}\|_m^2 \leq (1+\delta)|\mathbf{a}|^2, \quad a \in \mathbb{R}^n. \tag{3.116}$$

En particulier l'inégalité de droite signifie que

$$\|T\|_{2\to m} = \max\{\|T\mathbf{a}\|_m \ : \ |\mathbf{a}| = 1\} \leq (1+\delta)^{1/2}. \tag{3.117}$$

On a recours alors au lemme suivant.

Lemme 3.1 *Soit E et F des espaces de dimension finis munis de normes $\|\cdot\|_E$ et $\|\cdot\|_F$, Soit $0 < \delta_1 < 1$ et $0 < \delta_2 < 1$, et soit $R \subset E$ un ensemble fini tel qu'on ait la propriété de recouvrement*

$$\{\mathbf{x} \in E \ : \ \|\mathbf{x}\|_E = 1\} \subset \bigcup_{\mathbf{y}\in R} B(\mathbf{y}, \delta_1). \tag{3.118}$$

Soit $T : E \to F$ est un opérateur linéaire tel que

$$(1-\delta_2)\|\mathbf{y}\|_E^2 \leq \|T\mathbf{y}\|_F^2 \leq (1+\delta_2)\|\mathbf{y}\|_E^2, \quad \mathbf{y} \in R. \tag{3.119}$$

Alors on a aussi

$$A\|\mathbf{x}\|_E^2 \leq \|T\mathbf{x}\|_F^2 \leq B\|\mathbf{x}\|_E^2, \quad \mathbf{x} \in E. \tag{3.120}$$

avec $B = (1+\delta_2)\left(\frac{1+\delta_1}{1-\delta_1}\right)^2$ et $A = \left((1-\delta_2)^{1/2}(1-\delta_1) - (1+\delta_2)^{1/2}\frac{1+\delta_1}{1-\delta_1}\ \delta_1\right)^2$.

Preuve : on note

$$M := \|T\|_{E\to F} := \max\{\|T\mathbf{x}\|_F \ : \ \|\mathbf{x}\|_E = 1\}. \tag{3.121}$$

Pour n'importe quel $\mathbf{x} \in E$ tel que $\|\mathbf{x}\| = 1$, il existe $\mathbf{y} \in R$ tel que $\|\mathbf{x}-\mathbf{y}\|_E \leq \delta_1$. On peut donc écrire

$$\|T\mathbf{x}\|_F \leq \|T\mathbf{y}\|_F + \|T(\mathbf{x}-\mathbf{y})\|_F \leq (1+\delta_2)^{1/2}\|\mathbf{y}\|_E + M\delta_1 \leq (1+\delta_2)^{1/2}(1+\delta_1) + M\delta_1. \tag{3.122}$$

Ceci nous montre que

$$M \leq (1+\delta_2)^{1/2}(1+\delta_1) + M\delta_1, \tag{3.123}$$

et par conséquent $M \leq (1+\delta_2)^{1/2}\frac{1+\delta_1}{1-\delta_1}$ ce qui entraîne l'inégalité de droite dans (3.120). D'autre part on peut aussi écrire

$$\begin{aligned}\|T\mathbf{x}\|_F &\geq \|T\mathbf{y}\|_F - \|T(\mathbf{x}-\mathbf{y})\|_F \geq (1-\delta_2)^{1/2}\|\mathbf{y}\|_E - M\delta_1 \\ &\geq (1-\delta_2)^{1/2}(1-\delta_1) - (1+\delta_2)^{1/2}\frac{1+\delta_1}{1-\delta_1}\ \delta_1.\end{aligned}$$

Par homogénéité ceci entraîne l'inégalité de gauche dans (3.120). □

On voit que les constantes A et B tendent vers 1 quand $\delta_1, \delta_2 \to 0$. Par exemple le choix $\delta_1 = 1/20$ et $\delta_2 = 1/5$ nous assure $A > 1/2$ et $B < 3/2$. Ce lemme nous montre que pour vérifier l'équivalence (3.40) avec $\delta = \frac{1}{2}$ uniformément sur V_n, il suffit de la vérifier avec $\delta = \frac{1}{5}$ sur un ensemble fini S de fonctions de V_n dont le cardinal est celui

d'un recouvrement R de la sphère unité de $\mathbb{R}^n$ par des boules de taille 1/20. Pour un tel ensemble on a donc

$$E_{1/2} \subset \bigcap_{v\in S} E_{\frac{1}{5},v} \implies \Pr(E_{1/2}^c) \leq \#(S)\Pr(E_{\frac{1}{5},v}^c). \tag{3.124}$$

On a vu qu'on peut trouver un recouvrement de la boule unité de dimension n qui soit de cardinal inférieur à $(3/\delta_1)^n = 60^n$. Il vient ainsi

$$\Pr(E_{1/2}^c) \leq 60^n\, 2\exp\Big(-\frac{mc_{1/5}}{K_{n,w}}\Big). \tag{3.125}$$

Ceci nous montre que $\Pr(E_{1/2}^c) < \varepsilon$ dès qu'on est dans le régime d'échantillonage

$$m \geq \frac{1}{c_{1/5}} K_{n,w}(\ln(2/\varepsilon) + dn), \quad d = \ln(60). \tag{3.126}$$

En particulier on a $m \geq nK_{n,w} \geq n^2$, qui est un budget d'échantillonage sous-optimal. Nous allons voir qu'on peut faire mieux avec une analyse différente.

On revient pour cela à l'interprêtation de l'équivalence de norme en termes matriciels :

$$E_\delta = \{\|\mathbf{G} - \mathbf{I}\|_2 \leq \ \delta\}, \tag{3.127}$$

On rappelle que les coefficients de la matrice de Gramm $\mathbf{G}$ sont donnés par

$$G_{j,k} = \langle L_j, L_k\rangle_m = \frac{1}{m}\sum_{i=1}^m w(x^i)L_j(x^i)L_k(x^i), \tag{3.128}$$

ce qui nous montre que

$$\mathbf{G} = \frac{1}{m}\sum_{i=1}^m \mathbf{X}_i, \tag{3.129}$$

où les $\mathbf{X}_i$ sont des réalisations indépendantes d'une variable aléatoire *matricielle* définie par

$$\mathbf{X} = (w(x)L_j(x)L_k(x))_{j,k=1,\dots,n}, \tag{3.130}$$

lorsque x suit la loi de probabilité σ. On a $\mathbb{E}(\mathbf{X}) = \mathbf{I}$ et donc

$$\Pr(E_\delta^c) = \Pr\Big(\Big\|\frac{1}{m}\sum_{i=1}^m \mathbf{X}_i - \mathbb{E}(\mathbf{X})\Big\|_2 > \delta\Big). \tag{3.131}$$

On cherche donc une inégalité de concentration en norme spectrale pour les moyennes empiriques de la variable $\mathbf{X}$. Quelques remarques sur cette matrice aléatoire :

1. Elle est symmétrique et donc diagonalisable dans une base orthonormée avec valeurs propres réelles, et a la forme

   $$\mathbf{X} = w(x)\mathbf{x}\mathbf{x}^T, \quad \mathbf{x} = (L_1(x), \dots, L_n(x))^T, \tag{3.132}$$

   ce qui montre que c'est une matrice de rang 1.

2. La forme quadratique associée est

$$\mathbf{a} \mapsto \mathbf{a}^T \mathbf{X} \mathbf{a} = w(x)(\mathbf{a}^T \mathbf{x})^2, \tag{3.133}$$

ce qui montre que c'est une matrice positive.

3. Sa norme spectrale est donnée par

$$\|\mathbf{X}\|_2 = \lambda_{\max}(\mathbf{X}) = \max\{\mathbf{a}^T \mathbf{X} \mathbf{a} \ : \ |\mathbf{a}| = 1\} = w(x)|\mathbf{x}|^2 = w(x) \sum_{j=1}^{n} |L_j(x)|^2 = k_{n,w}(x), \tag{3.134}$$

ce qui nous montre que c'est une variable bornée au sens où

$$\|\mathbf{X}\|_2 \leq K_{n,w}, \tag{3.135}$$

presque sûrement.

On note que c'est la même borne que celle qu'on avait utilisée pour appliquer l'inégalité de Chernoff à la variable scalaire $X = \frac{w(x)|v(x)|^2}{\|v\|^2}$ lorsque v est une fonction de V_n. On va donc chercher à procéder avec les variables matricielles comme pour l'inégalité de Chernoff établie pour les variables scalaires.

Si $\mathbf{X}$ est une variable aléatoire matricielle symmétrique positive telle que $\mathbb{E}(\mathbf{X}) = \mathbf{I}$ et $\|\mathbf{X}\|_2 \leq K$ presque sûrement, on écrit

$$\Pr\Big(\Big\|\frac{1}{m}\sum_{i=1}^{m} \mathbf{X}_i - \mathbf{I}\Big\|_2 > \delta\Big) = \Pr\Big(\lambda_{\max}\Big(\frac{1}{m}\sum_{i=1}^{m} \mathbf{X}_i\Big) > 1+\delta\Big) + \Pr\Big(\lambda_{\min}\Big(\frac{1}{m}\sum_{i=1}^{m} \mathbf{X}_i\Big) < 1-\delta\Big), \tag{3.136}$$

et on cherche à borner les deux termes. Examinons le traitement du premier terme. Pour $t > 0$ on passe à l'exponentielle matricielle en écrivant

$$\Pr\Big(\lambda_{\max}\Big(\frac{1}{m}\sum_{i=1}^{m} \mathbf{X}_i\Big) > 1+\delta\Big) = \Pr\Big(\lambda_{\max}\Big(\exp\Big(t\sum_{i=1}^{m} \mathbf{X}_i\Big)\Big) > \exp(mt(1+\delta))\Big), \tag{3.137}$$

puis avec Markov on obtient

$$\Pr\Big(\lambda_{\max}\Big(\frac{1}{m}\sum_{i=1}^{m} \mathbf{X}_i\Big) > 1+\delta\Big) \leq \exp(-m(1+\delta)t)\mathbb{E}\Big(\exp\Big(t\lambda_{\max}\Big(\sum_{j=1}^{m} \mathbf{X}_j\Big)\Big)\Big) \tag{3.138}$$

Rappelons ici que l'exponentielle d'une matrice $\mathbf{M}$ peut se définir comme la série entière $\sum_{k\geq 0} \frac{1}{k!}\mathbf{M}^k$. Dans le cas des matrice symmétriques, il est plus simple d'y penser comme la matrice dont les valeurs propres sont les exponentielles de celles de $\mathbf{M}$ associées aux mêmes vecteurs propres, ce qui montre que $\exp(\mathbf{M})$ est alors symmétrique définie positive. On définit de la même façon $\ln(\mathbf{M})$ pour une matrice symmétrique définie positive et on a $\ln(\exp(\mathbf{M})) = \mathbf{M}$ pour toute matrice symmétrique.

Le traitement de l'espérance dans le membre de droite de l'inégalité obtenue est plus délicat que dans le cas scalaire pour lequel on avait simplement $\mathbb{E}(\exp(tX))^m$. Une approche immédiate serait d'écrire

$$\lambda_{\max}\Big(\sum_{j=1}^{m} \mathbf{X}_j\Big) = \Big\|\sum_{j=1}^{m} \mathbf{X}_j\Big\|_2 \leq \sum_{j=1}^{m} \|\mathbf{X}_j\|_2, \tag{3.139}$$

mais la dernière inégalité est beaucoup trop violente pour des matrices de rang 1 indépendantes (exemple : pour des projecteurs sur des vecteurs qui sont orthogonaux, la somme reste un projecteur donc garde une norme égale à 1). De fait, on obtiendrait ainsi une borne par $\mathbb{E}(\exp(t\|\mathbf{X}\|_2))^m$, mais cela menerait à des estimations beaucoup moins bonnes que dans le cas scalaire car bien que $\mathbb{E}(\mathbf{X}) = \mathbf{I}$ on a pas $\mathbb{E}(\|\mathbf{X}\|_2) = 1$ (la norme et l'espérance ne commutent pas). En fait on a vu que $\|\mathbf{X}\|_2 = k_{n,w}(x)$ et par conséquent

$$\mathbb{E}(\|\mathbf{X}\|_2) = \int_D k_{n,w}(x)d\sigma = n. \tag{3.140}$$

Une autre manière de traiter le problème est de borner la valeur propre maximale par la trace en écrivant

$$\exp\Big(t\lambda_{\max}\Big(\sum_{j=1}^m \mathbf{X}_j\Big)\Big) = \lambda_{\max}\Big(\exp\Big(t\sum_{j=1}^m \mathbf{X}_j\Big)\Big) \leq \mathrm{tr}\Big(\exp\Big(t\sum_{j=1}^m \mathbf{X}_j\Big)\Big), \tag{3.141}$$

ce qui parait violent mais va conduire à un meilleur résultat. La trace étant linéaire, on peut la faire commuter avec l'espérance et obtenir ainsi

$$\mathbb{E}\Big(\exp\Big(t\lambda_{\max}\Big(\sum_{j=1}^m \mathbf{X}_j\Big)\Big)\Big) \leq \mathrm{tr}\Big(\mathbb{E}\Big(\exp\Big(t\sum_{j=1}^m \mathbf{X}_j\Big)\Big)\Big). \tag{3.142}$$

On fait alors face à une nouvelle difficulté : on aimerait écrire $\exp\Big(t\sum_{j=1}^m \mathbf{X}_j\Big) = \prod_{j=1}^m \exp(t\mathbf{X}_j)$ ce qui permettrait de poursuivre et conclure de manière très similaire à l'inégalité de Chernoff pour les variables scalaires. Or ceci est uniquement vrai si les matrices commutent ce qui n'a aucune raison d'être le cas ici. La présence de la trace va cependant nous être utile car elle va permettre d'exploiter des résultats profonds sur les matrices que nous allons citer sans démonstration. Une première intuition provient de l'inégalité de Golden-Thompson qui affirme que si $\mathbf{M}_1$ et $\mathbf{M}_2$ sont des matrices symmétriques positives on a

$$\mathrm{tr}(\exp(\mathbf{M}_1 + \mathbf{M}_2)) \leq \mathrm{tr}(\exp(\mathbf{M}_1)\exp(\mathbf{M}_2)). \tag{3.143}$$

Malheureusement cette inégalité n'est plus vraie si on considère des sommes $\mathbf{M}_1 + \cdots + \mathbf{M}_m$ de plus de deux matrices, elle se détériore si on cherche à l'itérer. L'approche la plus efficace utilise un résultat difficile du à Lieb qui est la concavité de l'application

$$\mathbf{X} \mapsto \mathrm{tr}(\exp(\mathbf{M} + \ln(\mathbf{X}))), \tag{3.144}$$

sur le cône des matrices symmétrique positives, lorsque $\mathbf{M}$ est une matrice symmétrique fixée (notons que la propriété analogue dans le cas de variables scalaires est triviale puisque l'application devient alors $x \mapsto e^m x$). Ceci permet d'établir que si $\mathbf{X}$ est une matrice symmétrique aléatoire positive on a par concavité

$$\mathbb{E}(\mathrm{tr}(\exp(\mathbf{M} + \mathbf{X})) = \mathbb{E}(\mathrm{tr}(\exp(\mathbf{M} + \ln(\exp(\mathbf{X}))) \leq \mathrm{tr}(\exp(\mathbf{M} + \ln(\mathbb{E}(\exp(\mathbf{X})))). \tag{3.145}$$

De cette manière on peut écrire

$$\mathrm{tr}\Big(\mathbb{E}\Big(\exp\Big(t\sum_{j=1}^m \mathbf{X}_j\Big)\Big)\Big) = \mathrm{tr}\Big(\mathbb{E}_{m-1}\Big(\exp\Big(t\sum_{j=1}^{m-1} \mathbf{X}_j + \ln(\mathbb{E}(\exp(t\mathbf{X}_m)))\Big)\Big)\Big), \tag{3.146}$$

où $\mathbb{E}_{m-1}$ signifie que l'espérance ne porte plus que sur le tirage des variables $\mathbf{X}_1, \ldots, \mathbf{X}_{m-1}$. Par récurrence (**exercice**) on obtient ainsi

$$\mathrm{tr}\Big(\mathbb{E}\Big(\exp\Big(t\sum_{j=1}^{m}\mathbf{X}_j\Big)\Big)\Big) \leq \mathrm{tr}\Big(\exp\Big(\sum_{j=1}^{m}\ln(\mathbb{E}(\exp(t\mathbf{X}_j)))\Big)\Big) = \mathrm{tr}\Big(\mathbb{E}(\exp(t\mathbf{X}))^m\Big). \quad (3.147)$$

A partir de là on peut à nouveau raisonner essentiellement comme dans le cas scalaire en bornant au préalable la trace par

$$\mathrm{tr}\Big(\mathbb{E}(\exp(t\mathbf{X}))^m\Big) \leq n\lambda_{\max}\Big(\mathbb{E}(\exp(t\mathbf{X}))^m\Big) = n(\lambda_{\max}(\mathbb{E}(\exp(t\mathbf{X})))^m, \quad (3.148)$$

puis en estimant

$$\lambda_{\max}(\mathbb{E}(\exp(t\mathbf{X}))) \leq 1 + K^{-1}(\exp(tK) - 1), \quad (3.149)$$

et en choisissant la même valeur particulière $t = K^{-1}\ln(1+\delta) > 0$. Par un traitement similaire du terme $\Pr\Big(\lambda_{\min}\Big(\frac{1}{m}\sum_{i=1}^{m}\mathbf{X}_i\Big) < 1-\delta\Big)$, cela nous conduit à l'inégalité de Chernoff matricielle suivante.

Théorème 3.2 *Soit $\mathbf{X}_i$ des réalisations indépendantes d'une variable aléatoire matricielle $\mathbf{X}$ symmétrique positive, telle que $\mathbb{E}(\mathbf{X}) = \mathbf{I}$ et telle que $\|\mathbf{X}\|_2 \leq K$ presque sûrement. Alors, pour $0 < \delta < 1$ on a*

$$\Pr\Big(\Big\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{X}_i - \mathbf{I}\Big\|_2 > \delta\Big) \leq 2n\exp\Big(-\frac{c_\delta m}{K}\Big), \quad (3.150)$$

avec $c_\delta := (1+\delta)\ln(1+\delta) - \delta > 0$.

Un traitement détaillé des inégalités de concentration matricielles, qui inclut en particulier une version plus générale de l'inégalité de Chernoff ci-dessus, se trouve dans la référence [4] de Tropp citée en introduction. Pour les matrices qui nous intéressent cela donne

$$\Pr(E_\delta^c) \leq 2n\exp\Big(-\frac{c_\delta m}{K_{n,w}}\Big), \quad (3.151)$$

en en particulier $\Pr(E_\delta^c) \leq \varepsilon$ sous le régime d'échantillonage

$$m \geq c_\delta^{-1} K_{n,w}\ln(2n/\varepsilon). \quad (3.152)$$

On voit ainsi qu'au final, le passage d'une inégalité de concentration pour une seule fonction de V_n à une inégalité de concentration uniforme sur V_n n'est pas si coûteux : il se traduit par la présence supplémentaire d'un n à l'intérieur du logarithme.

3.5 Résultats de convergence

Nous pouvons à présent établir des estimations d'erreur pour l'estimateur des moindres carrés conditionné à l'évenement E_δ. On va systématiquement prendre la valeur $\delta = \frac{1}{2}$, et on considère donc

$$\tilde{u} = u_n\chi_{E_{1/2}}. \quad (3.153)$$

On étudie d'abord le cas où l'on observe des données non-bruités $y^i = u(x^i)$. Comme on l'a déjà remarqué,

$$\mathbb{E}(\|u - \tilde{u}\|^2) \leq \mathbb{E}(\|u - u_n\|^2 \chi_{E_{1/2}}) + \Pr(E_{1/2}^c)\|u\|^2. \tag{3.154}$$

Nous avons vu que sous le régime d'échantillonage

$$m \geq c^* K_{n,w} \ln(2n/\varepsilon), \quad c^* = \frac{1}{c_{1/2}} \approx 10, \tag{3.155}$$

on a $\Pr(E_{1/2}^c) \leq \varepsilon$ et on peut ainsi contrôler le deuxième terme en choisissant une valeur de ε correspondant à la précision souhaitée.

Nous allons évaluer le premier terme un peu plus finement que nous l'avons fait auparavant. On a vu que $u_n = P_n^m u$ est la projection de u sur V_n au sens de la norme $\|\cdot\|_m$. Ceci nous permet d'écrire

$$\|u - u_n\|^2 = \|u - P_n u\|^2 + \|P_n u - P_n^m u\|^2 = e_n(u)^2 + \|P_n^m(u - P_n u)\|^2, \tag{3.156}$$

puisque $P_n u = P_n^m P_n u$ par la propriété de projection. En posant

$$g = u - P_n u, \tag{3.157}$$

on a donc

$$\mathbb{E}(\|u - u_n\|^2 \chi_{E_{1/2}}) \leq e_n(u)^2 + \mathbb{E}(\|P_n^m g\|^2 \chi_{E_{1/2}}). \tag{3.158}$$

On cherche donc à évaluer $\|P_n^m g\|^2$ sous l'évènement $E_{1/2}$. On note que $P_n^m g$ est l'approximation des moindres carrés appliquée aux données $g(x^i)$. Si on la décompose suivant $P_n^m g = \sum_{j=1}^n c_j L_j$, on a

$$\|P_n^m g\|^2 = |\mathbf{c}|^2, \tag{3.159}$$

et le vecteur $\mathbf{c} = (c_1, \ldots, c_n)^T$ est solution de

$$\mathbf{G}\mathbf{c} = \mathbf{d}, \tag{3.160}$$

avec un second membre $\mathbf{d} = (d_1, \ldots, d_n)^T$ donné par

$$d_j = \frac{1}{m} \sum_{i=1}^m w(x^i) L_j(x^i) g(x^i). \tag{3.161}$$

Sous l'évenement $E_{1/2}$, on sait que l'on a $\lambda_{\min}(\mathbf{G}) \geq 1/2$ c'est à dire

$$\|\mathbf{G}^{-1}\|_2 \ \leq 2, \tag{3.162}$$

et par conséquence

$$|\mathbf{c}|2 \leq 4|\mathbf{d}|^2 = \frac{4}{m^2} \sum_{j=1}^n \Big| \sum_{i=1}^m w(x^i) L_j(x^i) g(x^i) \Big|^2. \tag{3.163}$$

On peut donc écrire

$$\mathbb{E}(\|P_n^m g\|^2 \chi_{E_{1/2}}) \leq \frac{4}{m^2} \sum_{j=1}^{n} \mathbb{E}\Big(\Big| \sum_{i=1}^{m} w(x^i) L_j(x^i) g(x^i) \Big|^2 \Big). \tag{3.164}$$

En notant T_j le j-ème terme de la somme ci-dessus, celui-ci peut se développer suivant

$$\begin{aligned} T_j &= \sum_{i=1}^{m} \mathbb{E}(|w(x^i) L_j(x^i) g(x^i)|^2) + \sum_{k \neq i} \mathbb{E}(w(x^i) L_j(x^i) g(x^i) w(x^k) L_j(x^k) g(x^k)) \Big) \\ &= m \int_D w(x)^2 |L_j(x)|^2 |g(x)|^2 d\sigma + m(m-1) \Big(\int_D w(x) L_j(x) g(x) d\sigma \Big)^2 \\ &= m \int_D w(x) |L_j(x)|^2 |g(x)|^2 d\mu + m(m-1) \Big(\int_D L_j(x) g(x) d\mu \Big)^2. \end{aligned}$$

Comme $g = u - P_n u$ est orthogonal aux L_j, on voit que le deuxième terme est nul, et par conséquent $T_j = m \int_D w(x) |L_j(x)|^2 |g(x)|^2 d\mu$. En sommant sur j on obtient ainsi

$$\mathbb{E}(\|P_n^m g\|^2 \chi_{E_{1/2}}) \leq \frac{4}{m} \int_D k_{n,w}(x) |g(x)|^2 d\mu \leq \frac{4K_{n,w}}{m} \|g\|^2 = \frac{4K_{n,w}}{m} e_n(u)^2. \tag{3.165}$$

Sous le régime d'échantillonage (3.155), la constante multiplicative ci-dessus est bornée suivant

$$\frac{4K_{n,w}}{m} \leq \delta(n,\varepsilon) := \frac{4}{c \ln(2n/\varepsilon)}. \tag{3.166}$$

et par conséquent,

$$\mathbb{E}(\|u - u_n\|^2 \chi_{E_{1/2}}) \leq (1 + \delta(n,\varepsilon)) e_n(u)^2. \tag{3.167}$$

On note que $\delta(n,\varepsilon) \to 0$ quand $n \to \infty$ ou $\varepsilon \to 0$. Nous avons donc obtenu le résultat suivant.

Théorème 3.3 *Sous le régime d'échantillonage* (3.155)*, pour toute fonction $u \in L^2(D,\mu)$, l'estimateur $\tilde{u}$ vérifie*

$$\mathbb{E}(\|u - \tilde{u}\|^2) \leq (1+\delta) e_n(u)^2 + \varepsilon \|u\|^2. \tag{3.168}$$

où $\delta = \delta(n,\varepsilon) \to 0$ quand $n \to \infty$ ou $\varepsilon \to 0$.

Notons que dans le cas particulier où on a choisi la mesure d'échantillonage

$$d\sigma = d\sigma^* := \frac{k_n}{n} d\mu, \tag{3.169}$$

et le poids correspondant

$$w = w^* = \frac{n}{k_n}, \tag{3.170}$$

on a vu qu'alors $K_{n,w^*} = n$ et par conséquent la condition sur le régime d'échantillonage devient alors

$$m \geq c^* n \ln(2n/\varepsilon). \tag{3.171}$$

On peut ainsi utiliser un budget d'échantillonnage presque optimal au sens où $m \sim n$ à des facteurs logarithmiques près. Pour cette raison, on peut considérer que la mesure

σ^* décrit un "échantillonage aléatoire optimal". Nous reviendrons un peu plus loin sur les propriétés de cette mesure.

Nous allons donner plusieurs variantes du résultat de convergence exprimé dans le Théorème 3.3. La première consiste à étudier la vitesse de convergence de l'estimateur lorsqu'on fait varier n. On suppose que la fonction u appartient à une classe $\mathcal{K}$ telle que, pour une certaine suite d'espace d'approximation $(V_n)_{n\geq 0}$ on a

$$e_n(\mathcal{K}) = \sup_{u\in\mathcal{K}} e_n(u) \leq Cn^{-s}, \quad n > 0. \tag{3.172}$$

On peut alors choisir d'utiliser ces espaces V_n pour construire une suite d'estimateurs des moindres carrés à poids. Il est alors naturel de choisir $\varepsilon = n^{-2s}$ pour équilibrer les deux termes dans l'estimation d'erreur, de sorte que l'on ait la même vitesse de convergence pour $\tilde{u} = \tilde{u}_n \in V_n$ au sens où

$$\mathbb{E}(\|u - \tilde{u}_n\|^2) \leq Bn^{-2s}, \quad B := 1 + \delta + C^2. \tag{3.173}$$

On note alors que le régime d'échantillonage prend la forme

$$m \geq cK_{n,w} \ln(2n^{2s+1}) = C_n K_{n,w}, \quad C_n = c(2s+1)\ln(2n), \tag{3.174}$$

c'est à dire $m \geq K_{n,w}$ à un facteur logarithmique en n près. Dans le cas où on utilise la mesure d'échantillonage σ^*, on obtient ainsi la vitesse d'approximation optimale avec un budget $m = \mathcal{O}(n\ln(n))$.

Une deuxième variante va nous permettre de nous affranchir de l'espérance dans l'estimation d'erreur en faisant intervenir l'erreur de meilleur approximation en norme L^∞,

$$e_n(u)_{L^\infty} = \min_{v\in V_n} \|u - v\|_{L^\infty}, \tag{3.175}$$

lorsqu'on suppose que $u \in L^\infty(D)$. On remarque que sous l'évenement $E_{1/2}$, l'estimateur des moindres carrés à poids u_n vérifie, pour toute fonction $v \in V_n$

$$\|u-u_n\| \leq \|u-v\| + \|u_n - v\| \leq \|u-v\| + \sqrt{2}\|u_n - v\|_m \leq \|u-v\| + \sqrt{2}\|u-v\|_m, \tag{3.176}$$

où on a utilisé le fait que u_n est la projection $P_n^m u$ dans la deuxième inégalité et qu'on peut appliquer le théorème de Pythagore pour la norme $\|\cdot\|_m$. Dans le cas où μ est une mesure de probabilité on sait que

$$\|v\| = \Big(\int_D |v(x)|^2 d\mu\Big)^{1/2} \leq \|v\|_{L^\infty}, \tag{3.177}$$

pour toute fonction $v \in L^\infty(D)$. Ainsi on peut borner le premier terme par

$$\|u - v\| \leq \|u - v\|_{L^\infty}. \tag{3.178}$$

Le deuxième terme peut lui aussi être borné par la norme L^∞ dans plusieurs circonstances : si on travaille avec les moindres carrés sans poids, c'est à dire $w = 1$, on a pour toute fonction $v \in L^\infty(D)$,

$$\|v\|_m = \Big(\frac{1}{m}\sum_{i=1}^m |v(x^i)|^2\Big)^{1/2} \leq \|v\|_{L^\infty}, \tag{3.179}$$

presque sûrement. Si on travaille avec des moindres carrés à poids, on peut aussi obtenir une telle estimation en supposant que les espace V_n utilisés contiennent les fonctions constantes. C'est le cas pour tous les espaces d'approximations classiquement utilisés : polynômes algébriques ou trigonométriques, fonctions constantes ou polynomiales par morceaux, espaces d'éléments finis. Dans ce cas, sous l'évenement $E_{1/2}$, on peut appliquer l'équivalence de norme à la fonction $1 \in V_n$ et obtenir ainsi

$$\frac{1}{m}\sum_{i=1}^{m} w(x^i) = \|1\|_m^2 \leq \frac{3}{2}\,\|1\|^2 = \frac{3}{2}\int_D d\mu = \frac{3}{2}. \tag{3.180}$$

Ceci entraîne que pour toute fonction $v \in L^\infty(D)$, on a

$$\|v\|_m = \Big(\frac{1}{m}\sum_{i=1}^{m} w(x_i)|v(x^i)|^2\Big)^{1/2} \leq \sqrt{3/2}\|v\|_{L^\infty}, \tag{3.181}$$

presque sûrement. On a ainsi obtenu

$$\|u - u_n\| \leq C\|u - v\|_{L^\infty}, \quad v \in V_n, \tag{3.182}$$

avec $C = 1 + \sqrt{3}$ presque sûrement sous l'évènement $E_{1/2}$. Notons qu'on peut enlever le "presque sûrement" en supposant que u est continue ainsi que les fonctions des espaces V_n, et on a ainsi obtenu le résultat suivant.

Théorème 3.4 *On suppose que μ est une mesure de probabilité, et que $w = 1$ ou que les fonctions constantes sont contenues dans l'espace V_n. Alors sous l'évenement $E_{1/2}$ - et donc avec probabilité supérieure à $1 - \varepsilon$ sous le régime d'échantillonage* (3.155) *- on a pour toute fonction $u \in \mathcal{C}(D)$*

$$\|u - u_n\| \leq Ce_n(u)_{L^\infty}, \tag{3.183}$$

avec $C = 1 + \sqrt{3}$.

On note que que si μ n'est pas une mesure de probabilité mais a une masse $\mu(D)$ finie, le résultat reste valable en multipliant la constante C par $\mu(D)^{1/2}$. L'intérêt de ce résultat est son caractère "uniforme" sur les fonctions u : on tire les x^i et si l'évènement $E_{1/2}$ est vérifié, l'estimation d'erreur est valable pour toute les fonctions u. En particulier, dans une stratégie numérique on peut choisir de répéter le tirage des x^i jusqu'à ce que $\|\mathbf{G} - \mathbf{I}\|_2 \leq \frac{1}{2}$, de façon à ce que l'estimation d'erreur (3.183) soit vérifié avec certitude pour toute fonction $u \in \mathcal{C}(D)$. L'inconvénient est que l'on majore l'erreur en norme L^2 par l'erreur de meilleure approximation en norme L^∞. Idéalement on aurait souhaité obtenir une erreur de la forme

$$\|u - u_n\| \leq Ce_n(u), \tag{3.184}$$

pour toute fonction $u \in L^2(D, \mu)$ sous l'évenement $E_{1/2}$. On peut cependant montrer qu'on ne peut pas espérer avoir une telle estimation : pour tout tirage $x^1, \ldots, x^m$ il existera toujours des fonctions $u \in L^2(D, \mu)$ telles $e_n(u)$ est arbitrairement petit mais $u - u_n$ reste grand (**exercice**).

Une troisième variante concerne le cas des données bruitées. On prend ici le modèle où

$$y^i = u(x^i) + \eta^i \tag{3.185}$$

où les η^i sont des variables centrées indépendantes et de variance $\eta^i = \kappa^2(x^i)$, où $x \mapsto \kappa(x)$ décrit le niveau du bruit au point x. En faisant la même analyse que précèdemment, on peut écrire

$$\|u - u_n\|^2 = e_n(u)^2 + \|P_n^m g + v\|^2 \leq e_n(u)^2 + 2\|P_n^m g\|^2 + 2\|v\|^2 \tag{3.186}$$

où $g = u - P_n u$ et $v = \sum_{j=1}^n \tilde{c}_j L_j$ est la fonction de V_n qui est obtenue en appliquant la méthode des moindres carrés sur les données de bruit $(\eta^1, \ldots, \eta^n)$ (on s'est servi ici de la linéarité de l'application $(y^1, \ldots, y^n) \to u_n$). On a déjà estimé $\mathbb{E}(\|P_n^m g\|^2 \chi_{E_{1/2}})$ et on va estimer par des techniques similaires la quantité $\mathbb{E}(\|v\|^2 \chi_{E_{1/2}})$. Le vecteur $\tilde{\mathbf{c}} = (\tilde{c}_1, \ldots, \tilde{c}_n)^T$ est solution de

$$\mathbf{G}\tilde{\mathbf{c}} = \tilde{\mathbf{d}} = (\tilde{d}_1, \ldots, \tilde{d}_n)^T, \quad \tilde{d}_j = \frac{1}{m} \sum_{i=1}^m w(x^i) L_j(x^i) \eta^i, \tag{3.187}$$

et on a ainsi

$$\mathbb{E}(\|h\|^2 \chi_{E_{1/2}}) \leq \frac{4}{m^2} \sum_{j=1}^n \mathbb{E}\Big(\Big| \sum_{i=1}^m w(x^i) L_j(x^i) \eta^i \Big|^2 \Big). \tag{3.188}$$

Noter que l'espérance porte à la fois sur le tirage des x^i et des η^i. On calcule chaque T_j en développant le carré à l'intérieur de l'espérance, on trouve à nouveau que les termes croisés s'annulent et on obtient finalement $T_j = m \int_D w(x) |L_j(x)|^2 \kappa(x)^2 d\mu$, ce qui conduit par sommation à une contribution du bruit de la forme

$$\mathbb{E}(\|v\|^2 \chi_{E_{1/2}}) \leq \frac{4}{m} \int_D k_{n,w}(x) \kappa(x)^2 d\mu. \tag{3.189}$$

Nous avons ainsi obtenu le résultat suivant.

Théorème 3.5 *Sous le régime d'échantillonage* (3.155) *et dans le cas de données bruitées, pour toute fonction $u \in L^2(D, \mu)$, l'estimateur $\tilde{u}$ vérifie*

$$\mathbb{E}(\|u - \tilde{u}\|^2) \leq (1 + 2\delta) e_n(u)^2 + \varepsilon \|u\|^2 + \frac{8}{m} \int_D k_{n,w}(x) \kappa(x)^2 d\mu. \tag{3.190}$$

où $\delta = \delta(n, \varepsilon) \to 0$ quand $n \to \infty$ ou $\varepsilon \to 0$.

On peut préciser le terme supplémentaire lié au bruit suivant les hypothèses de travail. Ainsi, si le bruit est uniforme $\kappa(x) = \kappa$ ou uniformément borné $\kappa(x) \leq \kappa$, on peut écrire

$$\int_D k_{n,w}(x) \kappa(x)^2 d\mu \ \leq \kappa^2 \int_D k_{n,w}(x) d\mu. \tag{3.191}$$

Dans le cas de la régression, on utilise $\sigma = \mu$ et $w = 1$, et on a ainsi $\int_D k_{n,w}(x) d\mu = \int_D k_n(x) d\mu = n$. Le terme de bruit est ainsi borné par $8 \frac{n}{m} \kappa^2$.

Dans le cas où on travaille avec la mesure d'échantillonage optimale $\sigma = \sigma^*$ et le poids correspondant $w = w^*$, on a $k_{n,w} = n$ et par conséquent

$$\int_D k_{n,w}(x)\kappa(x)^2 d\mu \ \leq n \int_D \kappa(x)^2 d\mu. \tag{3.192}$$

Le terme de bruit est ainsi borné par $\frac{n}{m}\int_D \kappa(x)^2 d\mu$. On note que si μ est une mesure de probabilité, on retrouve la borne $8\frac{n}{m}\kappa^2$ dans le cas d'un bruit uniforme ou borné par κ.

On voit ainsi que l'estimation fait apparaitre une compétition entre le terme d'erreur d'approximation $e_n(u)^2$ qui décroit avec n et ne dépend pas de m, et le terme de bruit $8\frac{n}{m}\kappa^2$ qui augmente avec n et décroit avec m. Ceci nous montre que pour un budget d'échantillonage m donné, il y aura une valeur optimale de n à trouver pour minimiser l'erreur. C'est l'illustration d'un *compromis biais-variance*, selon le language de la statistique mathématique. Si on suppose par exemple que u appartient à une classe $\mathcal{K}$ telle que l'on ait (3.172) c'est à dire une erreur d'approximation $e_n(u) = \mathcal{O}(n^{-s})$, et si m est dans le régime d'échantillonage (3.174), on voit alors que l'erreur d'estimation $\mathbb{E}(\|u - \tilde{u}\|^2)$ est bornée par une quantité de l'ordre de

$$n^{-2s} + \frac{n}{m}\kappa^2. \tag{3.193}$$

La valeur de n minimisant cette quantité sera obtenue en égalisant les deux termes, d'où $\kappa^2 n^{1+2s} = m$. En supposant que cette valeur de m est dans le régime d'échantillonage (3.174), on obtient ainsi une erreur d'estimation de la forme

$$\mathbb{E}(\|u - \tilde{u}\|^2) \ \leq C\Big(\frac{m}{\kappa^2}\Big)^{-r}, \quad r := \frac{2s}{1+2s}. \tag{3.194}$$

Noter que l'exposant r tend vers 1 lorsque $s \to \infty$, qui est la vitesse d'estimation d'une constante par m observations bruitées comme on l'avait remarqué dans le chapitre introductif.

3.6 Echantillonage optimal

Nous allons nous intéresser d'un peu plus près à la mesure d'échantillonage optimale

$$d\sigma^* = d\sigma_n^* = \frac{k_n}{n} d\mu, \tag{3.195}$$

dont on a vu qu'elle permet d'obtenir l'ensemble des résultats de convergence avec un budget

$$m = m(n, \varepsilon) = \lceil c^* n \log(2n/\varepsilon) \rceil, \qquad c^* = \frac{1}{c_{\frac{1}{2}}} \approx 10. \tag{3.196}$$

Une remarque importante est que σ_n^* dépend de V_n, et en particulier varie avec n lorsqu'on considère une suite d'espaces d'approximation $(V_n)_{n\geq 1}$.

Afin de comprendre l'intérêt de cette mesure, considérons les espaces $V_n = \mathbb{P}_{n-1}$ des polynômes de degré $n-1$ en dimension 1.

1. Si on travaille sur $D = [-1, 1]$ avec la mesure $d\mu = \frac{1}{\pi\sqrt{1-x^2}}$, on a vu que la base orthogonale est celle des polynômes de Chebychev pour laquelle après normalisation, on a $K_n = \|k_n\|_{L^\infty} = 2n+1$. Dans ce cas, les moindres carrés sans poids, qui correspondent à la mesure d'échantillonage $\sigma = \mu$ avec $w = 1$ et $K_{n,w} = K_n$, permettent d'obtenir l'ensemble des résultats de convergence avec un budget d'échantillonage $m = \lceil c^*(2n + 1)\log(2n/\varepsilon)\rceil$ qui est donc du même ordre que celui fourni par la mesure optimale.
2. Si on travaille toujours sur $D = [-1, 1]$ mais avec la mesure uniforme $d\mu = \frac{1}{2}dx$, on a vu que la base des polynômes de Legendre nous donne $K_n = n^2$. Dans ce cas, les moindres carrés sans poids, nécessitent un budget d'échantillonage quadratique $m = \lceil c^* n^2 \log(2n/\varepsilon)\rceil$. Il devient donc intéressant d'utiliser la mesure optimale qui permet un budget plus léger d'un facteur n. Il est possible de montrer dans ce cas que lorsque n tend vers $+\infty$, cette mesure tend en loi vers la mesure de Chebychev $\frac{1}{\pi\sqrt{1-x^2}}$. On peut même prouver une borne uniforme du type

$$d\sigma_n^* \leq \frac{B}{\pi\sqrt{1-x^2}}dx, \qquad n \geq 0. \tag{3.197}$$

   Ceci permet de montrer (**exercice**) qu'on peut aussi dans ce cas choisir

$$d\sigma = \frac{dx}{\pi\sqrt{1-x^2}} \quad \text{et} \quad w(x) = \frac{\pi}{2}\sqrt{1-x^2}, \tag{3.198}$$

   indépendamment de n, et que le budget d'échantillonage requis sera alors de nouveau de la forme $m = \lceil \tilde{c}n\log(2n/\varepsilon)\rceil$ avec une constante $\tilde{c}$ un peu plus grande que c^*.
3. Si on travaille sur $D = \mathbb{R}$ avec la mesure gaussienne $d\mu = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx$, la base orthonormale est celle des polynômes de Hermite $\{H_0, \ldots, H_{n-1}\}$ convenablement normalisés. Dans ce cas, la fonction $k_n(x)$ n'est pas bornée sur $\mathbb{R}$ (mis à part dans le cas trivial $n = 0$ puisque $H_0 = 1$). On a donc $K_n = \infty$ et si on veut utiliser les moindres carrés sans poids, la théorie ne nous fournit aucun régime d'échantillonage permettant d'avoir les résultats de convergence de l'estimateur. On peut faire une étude ad-hoc basée sur la restriction de la mesure gaussienne à un intervalle borné bien choisi (**exercice difficile**) afin d'établir qu'on peut avoir $\|\mathbf{G}-\mathbf{I}\|_2 \leq \frac{1}{2}$ avec grande probabilité sous un budget d'échantillonage $m \sim \exp(dn)$ pour une certaine constante d, ce qui est numériquement prohibitif. Dans ce cas, il est particulièrement judicieux d'avoir recours à la mesure optimale

$$d\sigma_n^* = \frac{1}{n}\Big(\sum_{j=0}^{n-1}|H_j(x)|^2\Big)d\mu, \tag{3.199}$$

   qui permet de rétablir un budget linéaire en n au facteur logarithmique près.

Ces contraintes sur les régimes d'échantillonage peut être vérifiées numériquement, par exemple en examinant le conditionnement moyen de la matrice $\mathbf{G}$ en fonction de n et m dans les différents cas que nous venons de citer.

Un cas simple, mais intéressant pour saisir la nature de la mesure optimale est celui où l'on travaille sur un domaine $D \subset \mathbb{R}^d$ muni d'une mesure $d\mu$, par exemple la mesure de Lebesgue et que l'on considère l'espace V_n des fonctions constantes par morceaux sur une partition $\{D_1, \ldots, D_n\}$. La base orthonormée est donnée par les fonctions indicatrices

$$L_j = |D_j|^{-1/2} \chi_{D_j}, \quad j = 1, \ldots, n, \quad |D_j| = \mu(D_j), \tag{3.200}$$

ce qui nous montre que la mesure optimale a la forme

$$d\sigma_n^* = h_n(x) d\mu, \quad h_n(x) = \frac{1}{n} \sum_{j=1}^{n} |D_j|^{-1} \chi_{D_j}. \tag{3.201}$$

La fonction $h_n(x)$, parfois appelé *fonction de maillage*, décrit ainsi la densité locale de la partition puisqu'elle est inversement proportionnelle à la taille de l'élément dans lequel on se trouve. On voit ainsi qu'un tirage suivant la mesure σ_n^* a une probabilité identique, égale à $1/n$, de tomber dans chaque élément D_i. On peut aussi montrer (**exercice**) échantillonage de $m = m(n, \varepsilon)$ points suivant cette mesure aura avec grande probabilité la propriété que chaque élément D_j contient au moins un point x^i.

Le tirage de points indépendants suivant une mesure σ donnée est en soit un problème qui peut devenir numériquement délicat. Les ordinateurs permettent aisément de faire des tirages indépendants suivant des mesures simples, par exemple une loi uniforme sur un intervalle I borné. En tirant indépendamment d composantes $(x_1, \ldots, x_d)$ avec x_i suivant la loi uniforme sur I, on obtient le tirage d'une variable de loi uniforme sur le cube $Q = I^d$. Enfin, si $D \subset \mathbb{R}^d$ est contenu dans un tel cube Q, le tirage d'une variable uniforme sur D correspond à tirer x uniforme sur Q et effectuer un rejet si $x \notin D$.

Si σ admet une densité $d\sigma = p(x) d\mu$ où μ est la mesure uniforme sur D et la densité $x \mapsto p(x)$ est connue, il existe plusieurs stratégies pour tirer x suivant σ. En dimension $d = 1$, sur un intervalle $I = [a, b]$ on peut calculer la fonction de répartition

$$P(x) = \int_a^x p(t) d\mu(x), \tag{3.202}$$

et on peut alors prendre $x = P^{-1}(y)$ où y est une variable uniforme sur $[0, 1]$. Une méthode plus générale qui fonctionne en toute dimension est la méthode de rejet si l'on connait une borne B telle que

$$p(x) \leq B, \quad x \in D. \tag{3.203}$$

La méthode consiste alors à tirer x suivant μ et une variable y uniforme sur $[0, 1]$, puis d'accepter x si et seulement $y \leq \frac{p(x)}{B}$. Il convient de noter que plus B est grand, plus on peut s'attendre à avoir un rejet fréquent, et il faudra typiquement de l'ordre de Bm tirage de x pour obtenir m échantillons non-rejetés donc tirés suivant σ. Dans le cas qui nous intéresse, si μ est la mesure uniforme et si on veut utiliser la méthode de rejet pour tirer selon σ_n^*, la borne sera

$$B = B_n = \frac{K_n}{n}, \tag{3.204}$$

qui peut être grand. Il faut cependant noter que dans les applications où le coût de calcul est dominé par les évaluations de la fonction u, celles-ci ne sont effectuées qu'aux m points retenus au final.

Le fait que la mesure optimale σ_n^* dépende de V_n représente une difficulté lorsqu'on est dans la situation où l'on parcours une suite d'espaces $(V_n)_{n\geq 1}$ vérifiant la propriété d'emboitement

$$V_n \subset V_{n+1}. \tag{3.205}$$

Celle-ci peut être imposée à l'avance (par exemple les espaces de polynômes de degré $n-1$, ou fabriquée par un procédé adaptatif où V_{n+1} dépend de l'approximation u_n calculée dans l'espace V_n (raffinement de maillage, sélection d'une nouvelle fonction de base). Dans les deux cas on est confronté à une difficulté si on souhaite une méthode d'échantillonage hierarchique : l'échantillon $S_n := \{x^1, \dots, x^{m(n,\varepsilon)}\}$ tiré suivant la mesure σ_n^* pour calculer l'estimateur $u_n \in V_n$ n'obéit pas à la loi σ_{n+1}^* que l'on souhaiterait utiliser pour calculer le nouvel estimateur $u_{n+1} \in V_{n+1}$. Cela nous oblige donc a-priori à tirer un nouvel échantillon S_{n+1} de taille $m(n+1,\varepsilon)$ et nous empêche ainsi de recycler les échantillons précédent. Au bout de n étapes, on aura utilisé un budget total d'échantillonage

$$M(n,\varepsilon) = m(1,\varepsilon) + m(2,\varepsilon) + \cdots + m(n,\varepsilon) \geq \mathcal{O}(n^2), \tag{3.206}$$

qui est fortement sous-optimal. On peut en fait tirer partie de la remarque suivante : la loi σ_{n+1}^* a la structure de mélange

$$\sigma_{n+1}^* = \frac{n}{n+1}\sigma_n^* + \frac{1}{n+1}\rho_{n+1}, \tag{3.207}$$

où $d\rho_{n+1} = |L_{n+1}|^2 d\mu$ est aussi une loi de probabilité. De facon générale un mélange de lois $\{\alpha_1, \dots, \alpha_k\}$ a la forme d'une combinaison barycentrique

$$d\alpha = p_1 d\alpha_1 + \cdot + p_k d\alpha_k, \quad p_i \geq 0, \quad \sum_{i=1}^{k} p_i = 1. \tag{3.208}$$

et le tirage d'une variable suivant la loi α est équivalent à tirer suivant α_i avec probabilité p_i. Ceci nous montre qu'on peut fabriquer l'échantillon S_{n+1} en tirant x^i avec probabilité $1/(n+1)$ suivant la loi ρ_{n+1} et en recyclant avec probabilité $(1-1/(n+1))$ un point de S_n. On peut étudier plus finement une telle approche (**exercice difficile**) et montrer que le budget cumulé d'échantillonage $M(n)$ à l'étape n reste en moyenne proportionel à n à un facteur logarithmique près.

Pour terminer ce chapitre, faisons quelques remarques critiques sur les résultats de convergence que nous avons établi. Ceux-ci présentent deux faiblesses par rapport à l'objectif idéal de combiner une approximation optimale et un budget optimal :

1. D'une part l'échantillonage est aléatoire et les résultats ne sont ainsi pas valable de manière certaine, mais seulement dans un sens probabiliste : estimation de l'espérance $\mathbb{E}(\|u-u_n\|^2\chi_{E_{1/2}})$ ou estimation de $\|u-u_n\|$ avec valable avec grande probabilité. Il serait en particulier intéressant de disposer de méthodes d'échantillonage déterministe donnant des estimations toujours vraies.
2. D'autre part, le budget n'est pas tout à fait de l'ordre optimal $m \sim n$, puisque même avec l'utilisation de la mesure σ_n^*, on a un facteur multiplicatif logarithmique $c\ln(2n/\varepsilon)$.

On peut viser certaines améliorations, au vu d'un résultat profond obtenu par Batson, Spielman et Srivastava en 2014, que nous citons ici sans démonstration.

Théorème 3.6 *Soit $M \geq n$ et soit $(\mathbf{x}_i)_{i=1,\dots,M}$ une suite de vecteurs de $\mathbb{R}^n$ tels que*

$$\sum_{i=1}^{M} \mathbf{x}_i \mathbf{x}_i^t = \mathbf{I}. \tag{3.209}$$

Alors pour tout entier $d > 1$ il existe des poids $s_i \geq 0$ tels que

$$m = \#\{i \ : \ s_i \neq 0\} \leq dn, \tag{3.210}$$

et

$$\mathbf{I} \leq \sum_{i=1}^{M} s_i \mathbf{x}_i \mathbf{x}_i^t \leq \frac{d+1+2\sqrt{d}}{d+1-2\sqrt{d}} \mathbf{I}, \tag{3.211}$$

au sens des matrices symmétriques.

La preuve de ce résultat est constructive et donne un algorithme déterministe pour calculer les poids s_i. Une conséquence (**exercice**) est que si pour un $0 < \alpha < 1$, on a

$$(1-\alpha)\mathbf{I} \leq \sum_{i=1}^{M} \mathbf{x}_i \mathbf{x}_i^t \ \leq (1+\alpha)\mathbf{I} \tag{3.212}$$

alors il existe un sous ensemble $S \subset \{1, \dots, M\}$ de cardinal $m \leq 2n$ et des poids $s_i \geq 0$ tels que

$$(1-\delta)\mathbf{I} \leq \sum_{i \in S} s_i \mathbf{x}_i \mathbf{x}_i^t \ \leq (1+\delta)\mathbf{I}, \tag{3.213}$$

pour une valeur $0 < \delta < 1$ qui peut se calculer en fonction de α.

On peut appliquer ceci à notre problème en considérant les vecteurs

$$\mathbf{x}_i = \frac{1}{\sqrt{M}} (L_1(x^i), \dots, L_n(x^i))^T, \tag{3.214}$$

et en partant d'une grille de points $(x^i)_{i=1,\dots,M}$ déterministe suffisament fine pour être assuré de la propriété (3.212). Par exemple si μ est la mesure de probabilité uniforme, on voit qu'on peut prendre les x^i répartis suivant une grille uniforme dans D et on est alors assuré que

$$\lim_{M \to \infty} \ \frac{1}{M} \sum_{i=1}^{M} L_j(x^i) L_k(x^i) = \int_D L_j(x) L_k(x) d\mu = \delta_{j,k}, \tag{3.215}$$

par convergence des sommes de Riemann (en supposant par exemple que les fonctions de V_n sont continues par morceaux). Cela montre que (3.212) est vérifié pour M suffisament grand avec $\alpha > 0$ arbitrairement petit. La valeur de M peut ici être très grande par rapport à n, mais le résultat ci-dessus nous assure de pouvoir extraire par un algorithme déterministe un ensemble de $m \leq 2n$ points, qu'on peut noter à nouveau $x^1, \dots, x^m$ et des poids $w(x^i)$ tels que

$$(1-\delta)\mathbf{I} \leq \mathbf{G} \ \leq (1+\delta)\mathbf{I}, \tag{3.216}$$

où $\mathbf{G} = (\frac{1}{m}\sum_{i=1}^m w(x^i)L_j(x^i)L_k(x^i))_{j,k=1,\dots,n}$ est la matrice de Gramm de la méthode des moindres carrés à poids. On a ainsi construit de façon déterministe un échantillonage de $m \leq 2n$ points tels que l'évenement E_δ est réalisé de façon certaine.

En suivant le même raisonnement qui nous a conduit au Théorème 3.4 de la section précèdente, on obtient que si μ est une mesure de probabilité et si les fonctions constantes sont contenues dans V_n, la méthode des moindres carrés à poids réalisée avec cet échantillonage à $m = 2n$ points déterministes nous donne pour toute fonction $u \in \mathcal{C}(D)$

$$\|u - u_n\| \leq Ce_n(u)_\infty, \quad C = 1 + \sqrt{\frac{1+\delta}{1-\delta}}. \tag{3.217}$$

On a déjà remarqué que quelque soit la taille de l'échantillon m, il est impossible d'espérer un résultat de la forme $\|u-u_n\| \leq Ce_n(u)$ pour toute fonction $u \in L^2(D,\mu)$. Un probleme ouvert est le suivant : existe-t-il un échantillonage aléatoire de taille $m = \mathcal{O}(n)$ tel qu'on soit assuré d'avoir pour toute fonction $u \in L^2(D,\mu)$

$$\mathbb{E}(\|u - u_n\|^2) \leq Ce_n(u)^2. \tag{3.218}$$

4 Espaces linéaires optimaux

Le chapitre précèdent nous a permis d'obtenir des bornes presques optimales au sens où l'erreur d'estimation est controlée (en un sens probabiliste) par $e_n(u) = \min_{v\in V_n}\|u - v\|$. Dans ce chapitre on va se poser la question de trouver l'espace $V_n = \overline{V}_n$ permettant de rendre $e_n(u)$ la plus petite possible. On se place ici dans un espace de Hilbert V séparable général (qui pourrait par exemple être $L^2(D,\mu)$ ou un espace de Sobolev $H^s(D)$), et on pose

$$e_n(u) = e_n(u)_V = \min_{v\in V_n}\|u - v\| = \|u - P_{V_n}u\|, \tag{4.1}$$

où $\|\cdot\| := \|\cdot\|_V$. Bien entendu le problème de rendre $e_n(u)$ minimal n'a pas de sens si on considère une seule fonction u car il suffirait de prendre un espace V_n qui contient u pour annuler $e_n(u)$. On a vu que ce problème prend un sens précis si on considère une classe définie par un ensemble compact $\mathcal{K} \subset V$ pour laquel l'espace optimal $\overline{V}_n$ serait celui qui atteint le minimum dans la définition de l'épaisseur de Kolmogorov

$$d_n = d_n(\mathcal{K})_V = \inf_{\dim(V_n)=n}\max_{u\in\mathcal{K}}\|u - P_{V_n}u\|. \tag{4.2}$$

Un tel espace optimal est généralement difficile à calculer (lorsqu'il existe) et nous allons nous pencher dans ce chapitre sur une stratégie numérique qui permet de s'en approcher : la méthode des bases réduites. On trouvera plus de détails sur l'analyse de cette méthode dans la référence [3] de Cohen-DeVore citée en introduction.

Auparavant on va considérer un autre cadre dans lequel u est une variable aléatoire à valeur dans l'espace V et où on cherche l'espace $\overline{V}_n$ optimal pour la moyenne quadratique de l'erreur de projection

$$\sigma_n^2 = \sigma_n^2(u)_V := \min_{\dim(V_n)=n}\mathbb{E}(\|u - P_{V_n}u\|^2). \tag{4.3}$$

On verra que ce problème admet une solution simple donnée par *l'analyse en composante principale* (ACP). On rappelle que si la variable aléatoire u est à support dans $\mathcal{K}$ on a $\sigma_n \leq d_n$.

Un cadre typique d'application est celui où on considère la solution $u \in V$ d'une EDP

$$\mathcal{P}(u, y) = 0, \tag{4.4}$$

décrivant un phénomène physique, qui dépend de d paramètres $y = (y_1, \ldots, y_d) \in Y \subset \mathbb{R}^d$: pour chaque valeur de $y \in Y$ on a une solution $u \in V$, ce qui définit une application

$$y \in Y \mapsto u(y) \in V. \tag{4.5}$$

Dans le cadre déterministe, on s'intéresse à l'ensemble des solutions lorsque les paramètres varient, en considérant la classe

$$\mathcal{K} = \{u(y) \ : \ y \in Y\} \subset V. \tag{4.6}$$

On cherche alors des espaces V_n qui approchent au mieux l'ensemble des fonctions u de cette classe. Dans le cadre stochastique, les paramètres sont aléatoires avec une distribution de probabilité sur Y et la solution u devient ainsi une variable aléatoire dans V. On cherche alors des espaces V_n qui approchent au mieux u au sens de la moyenne quadratique ci-dessus.

4.1 Approximation en composantes principales

Si V est un espace de dimension finie, par exemple $V = \mathbb{R}^N$, on rappelle qu'un vecteur aléatoire est donné par

$$\mathbf{u} = (u_1, \ldots, u_N)^T, \tag{4.7}$$

où les coordonnées u_i sont des variables aléatoires scalaires. L'espérance est donnée par

$$\overline{\mathbf{u}} = \mathbb{E}(\mathbf{u}) = (\overline{u}_1, \ldots, \overline{u}_N) = (\mathbb{E}(u_1), \ldots, \mathbb{E}(u_n))^T, \tag{4.8}$$

en ayant supposé que les variables u_i ont des moments d'ordre 1 finis, c'est à dire $\mathbb{E}(|u_i|) < \infty$. Si on suppose que les u_i ont aussi des moments finis d'ordre 2, c'est à dire $\mathbb{E}(|u_i|^2) < \infty$, on peut définir la matrice de corrélation

$$\mathbf{R} = (R_{i,j})_{i,j=1,\ldots,n} \quad R_{i,j} = \mathbb{E}(u_i u_j), \tag{4.9}$$

ainsi que la matrice de covariance

$$\mathbf{T} = (T_{i,j})_{i,j=1,\ldots,n} \quad T_{i,j} = \mathbb{E}((u_i - \overline{u}_i)(u_j - \overline{u}_j)). \tag{4.10}$$

Notons que la diagonale de $\mathbf{T}$ correspond aux variances des u_j et que $\mathbf{T} = \mathbf{R}$ dans le cas où $\mathbf{u}$ est une variable centrée c'est à dire $\mathbb{E}(\mathbf{u}) = 0$.

Lorsqu'on travaille en dimension infinie, par exemple pour des espaces de Hilbert de fonctions tels que $V = L^2(D, \mu)$ ou $V = H^s(D)$, on considère une fonction aléatoire u

(ou processus aléatoire) que l'on peut identifier via une base orthonormée $(\psi_j)_{j\geq 1}$ à un vecteur aléatoire de dimension infinie

$$\mathbf{u} = (u_j)_{j\geq 1}, \qquad u = \sum_{j\geq 1} u_j \psi_j. \tag{4.11}$$

On suppose

$$\mathbb{E}(\|u\|^2) = \mathbb{E}(|\mathbf{u}|^2) = \sum_{j\geq 1} \mathbb{E}(|u_j|^2) < \infty, \tag{4.12}$$

ce qui nous permet de définir la matrice de corrélation $\mathbf{R} = (\mathbb{E}(u_i u_j))_{i,j\geq 1}$ et de même la matrice de covariance $\mathbf{T}$ qui coincide avec $\mathbf{R}$ dans le cas d'un processus centré $\mathbb{E}(u) = 0$.

L'analyse en composante principale va nous permettre d'identifier l'espace $\overline{V}_n$ qui atteint le minimum dans (4.3) à travers l'étude spectrale de la matrice $\mathbf{R}$. Notons tout d'abord que cette matrice peut-être vue comme la discrétisation dans la base $(\psi_j)_{j\geq 1}$ de l'opérateur de corrélation

$$R : v \mapsto \mathbb{E}(\langle u, v\rangle u), \tag{4.13}$$

puisque $\langle R\psi_i, \psi_j\rangle = R_{i,j}$. Cet opérateur est autoadjoint et positif puisque

$$\langle Rv, v\rangle = \mathbb{E}(|\langle u, v\rangle|^2). \tag{4.14}$$

C'est aussi un opérateur compact car il a la propriété de Hilbert-Schmidt

$$\sum_{i,j\geq 1} |R_{i,j}|^2 = \sum_{i,j\geq 1} \mathbb{E}(u_i u_j)^2 \leq \sum_{i,j\geq 1} \mathbb{E}(|u_i|^2)\mathbb{E}(|u_j|^2) = \mathbb{E}(\|u\|^2)^2 < \infty, \tag{4.15}$$

On peut donc le diagonaliser dans une base orthonormée $(\varphi_j)_{j\geq 1}$ associée à une suite de valeurs propres

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq \cdots \geq 0, \quad \lim_{k\to\infty} \lambda_k = 0. \tag{4.16}$$

Notons aussi qu'il s'agit de plus d'un opérateur à trace puisque

$$\mathrm{tr}(R) = \sum_{k\geq 1} \lambda_k = \sum_{j\geq 1} R_{j,j} = \sum_{j\geq 1} \mathbb{E}(|u_j|^2) = \mathbb{E}(\|u\|^2) < \infty. \tag{4.17}$$

La base des $(\varphi_j)_{j\geq 0}$ est appelée base de Karhunen-Loeve. Elle a la propriété de décorréler le processus u puisque si on le décompose dans cette base suivant $u = \sum_{j\geq 1} \tilde{u}_j \varphi_j$ on voit que

$$\mathbb{E}(\tilde{u}_i \tilde{u}_j) = \mathbb{E}(\langle u, \varphi_i\rangle\langle u, \varphi_j\rangle) = \langle R\varphi_i, \varphi_j\rangle = \lambda_j \delta_{i,j}. \tag{4.18}$$

Cette base nous fournit la réponse au problème du meilleur espace $\overline{V}_n$ comme le montre le résultat suivant.

Théorème 4.1 *L'espace* $\overline{V}_n = \mathrm{vect}\{\varphi_1, \ldots, \varphi_n\}$ *vérifie*

$$\min_{\dim(V_n)=n} \mathbb{E}(\|u - P_{V_n} u\|^2) = \mathbb{E}(\|u - P_{\overline{V}_n} u\|^2) = \sum_{k>n} \lambda_k. \tag{4.19}$$

Preuve : Remarquons tout d'abord qu'on a en effet

$$\mathbb{E}(\|u - P_{\overline{V}_n} u\|^2) = \sum_{k>n} \mathbb{E}(|\langle u, \varphi_k \rangle|^2) = \sum_{k>n} \lambda_k. \tag{4.20}$$

Notre objectif est de montrer que $\mathbb{E}(\|u - P_{V_n} u\|^2)$ est minimisé par $V_n = \overline{V}_n$, ce qui par Pythagore revient à montrer que $\mathbb{E}(\|P_{V_n} u\|^2)$ est maximisé par ce même choix. On procède par récurrence sur n. Pour $n = 1$, si $V_1 = \mathbb{R}\varphi$ avec $\|\varphi\| = 1$, on peut écrire

$$\mathbb{E}(\|P_{V_1} u\|^2 = \mathbb{E}(|\langle u, \varphi \rangle|^2) = \langle R\varphi, \varphi \rangle, \tag{4.21}$$

et la théorie spectrale nous indique que cette quantité est maximisée par le choix $\varphi = \varphi_1$ qui correspond à l'espace $\overline{V}_1$. Supposons la propriété vraie au rang n et soit V_{n+1} n'importe quel espace de dimension $n + 1$. Il existe un $\psi \in V$ de norme 1 contenu dans $V_{n+1} \cap \overline{V}_n^{\perp}$. Cet élément a donc la forme

$$\psi = \sum_{k>n} c_k \varphi_k, \quad \sum_{k>n} |c_k|^2 = 1, \tag{4.22}$$

et on peut écrire $V_{n+1} = V_n \oplus^{\perp} \mathbb{R}\psi$. Ainsi

$$\mathbb{E}(\|P_{V_{n+1}} u\|^2) = \mathbb{E}(\|P_{V_n} u\|^2) + \mathbb{E}(|\langle u, \psi \rangle|^2), \tag{4.23}$$

et on a d'une part par l'hypothèse de récurrence,

$$\mathbb{E}(\|P_{V_n} u\|^2) \leq \mathbb{E}(\|P_{\overline{V}_n} u\|^2) = \lambda_1 + \cdots + \lambda_n, \tag{4.24}$$

et d'autre part

$$\mathbb{E}(|\langle u, \psi \rangle|^2) = \sum_{k>n} |c_k|^2 \mathbb{E}(|\langle u, \varphi_k \rangle|^2) = \sum_{k>n} |c_k|^2 \lambda_k \leq \lambda_{n+1} \sum_{k>n} |c_k|^2 = \lambda_{n+1}. \tag{4.25}$$

Par conséquent, on obtient

$$\mathbb{E}(\|P_{V_{n+1}} u\|^2) \leq \lambda_1 + \cdots + \lambda_{n+1} = \mathbb{E}(\|P_{\overline{V}_{n+1}} u\|^2), \tag{4.26}$$

ce qui prouve la propriété au rang $n + 1$. □.

Remarque 4.1 *Dans la pratique statistique de l'ACP sur un vecteur aléatoire, on commence toujours par recentrer u, et on cherche a résoudre le problème de minimisation*

$$\min_{\dim(V_n)=n} \mathbb{E}(\|u - \overline{u} - P_{V_n}(u - \overline{u})\|^2). \tag{4.27}$$

c'est à dire approcher au mieux u par un espace affine de la forme $\overline{u} + V_n$. Les espaces optimaux $\overline{V}_n$ sont alors calculés de la même manière en remplacant $\mathbf{R}$ *par la matrice de covariance* $\mathbf{T}$.

Remarque 4.2 *La diagonalisation de l'opérateur R est équivalente à celle de la matrice* $\mathbf{R}$ *qui est de taille infinie lorsque l'on travaille dans un espace V de dimension infinie. Le calcul pratique des λ_k et φ_k nécessite par conséquent des approximations numériques. Typiquement, on est amené à se ramener à un espace de dimension $N < \infty$ par un procédé de discrétisation. Par exemple lorsque que $u = u(y) \in V$ est solution d'une EDP dependant d'un vecteur aléatoire $y \in Y \subset \mathbb{R}^d$ de paramètres physiques, on peut on peut considérer l'approximation numérique $u_h = u_h(y)$ de cette solution dans un espace d'élément fini V_h. La dimension $N = \dim(V_h)$ est alors typiquement le nombre de sommets du maillage. Notons que $u_h \in V_h$ est elle aussi aléatoire. On se ramène ainsi à l'étude spectrale d'un opérateur de corrélation qui agit sur V_h suivant*

$$v_h \mapsto Rv_h = \mathbb{E}(\langle u_h, v_h\rangle u_h) \tag{4.28}$$

ou de manière équivalente à l'étude de la matrice $N \times N$

$$R_{i,j} = \mathbb{E}(\langle u_h, \psi_i\rangle\langle u_h, \psi_j\rangle), \tag{4.29}$$

où $\{\psi_1, \ldots, \psi_N\}$ est une base de V_h. Les espaces $\overline{V}_n$ obtenus sont ainsi des sous-espaces de V_h et l'objectif est que ceux-ci approchent la solution exacte avec une précision du même ordre mais avec $n << N$.

Remarque 4.3 *Une autre approximation souvent nécessaire est celle de l'espérance qui n'est pas calculable de manière exacte. Dans l'exemple de la remarque précèdent, une stratégie d'approximation consiste à tirer un grand nombre solutions numériques aléatoires $u_h^1, \ldots, u_h^M$ indépendantes (en effectuant des tirages indépendants $y^1, \ldots, y^M \in Y$ et en calculant numériquement les $u_h^i = u_h(y^i) \in V_h$). Ces "exemples" de solutions sont parfois appellées "snapshots". On approche alors l'opérateur de covariance par sa version empirique*

$$v_h \mapsto \widetilde{R}v_h = \frac{1}{M}\sum_{i=1}^{M}\langle u_h^i, v_h\rangle u_h^i, \tag{4.30}$$

et l'analyse spectrale de celui-ci conduit à des espaces $\widetilde{V}_n \subset V_h$ qui approchent les espaces optimaux $\overline{V}_n$ avec une précision qu'il est possible de quantifier par une analyse plus poussée.

Remarque 4.4 *Dans le cas particulier de l'espace $V = L^2(D)$ pour la mesure de Lebesgue, l'opérateur de covariance R a la forme d'un opérateur intégral : pour $x \in D$, on a*

$$Rv(x) = \mathbb{E}(\langle u, v\rangle u(x)) = \mathbb{E}\Big(u(x)\int_D u(y)v(y)dy\Big) = \int_D K_u(x,y)v(y)dy, \tag{4.31}$$

où

$$K_u(x,y) = \mathbb{E}(u(x)u(y)), \quad x, y \in D, \tag{4.32}$$

est le noyau de corrélation du processus u. Un cas important, souvent rencontré en pratique, est celui des processus stationnaires pour lesquels cette fonction ne dépend que de la distance entre les points x et y c'est à dire

$$K_u(x,y) = k_u(x-y). \tag{4.33}$$

L'opérateur à R prend alors la forme d'une convolution $Rv(x) = \int_D k_u(x-y)v(y)dy$. *Dans le cas dit "cyclo-stationnaire" où* D *est par exemple le tore* $[0,1]^d$ *périodique, la fonction* k_u *est une fonction de période* 1 *en chaque variable et la base de Karhunen-Loeve coïncide alors avec celle des séries de Fourier* $e_k(x) = \exp(i2\pi\langle k, x\rangle)$ *pour* $k \in \mathbb{Z}^d$ *puisque celle-ci diagonalise les opérateurs de convolution périodiques (***exercice***). De façon générale, les bases de Karhunen-Loeve des processus stationnaires sont le plus souvent à support global sur le domaine* D.

4.2 Bases réduites et algorithmes greedy

La méthode des bases réduites a été développée spécifiquement dans le contexte d'EDP

$$\mathcal{P}(u, y) = 0. \tag{4.34}$$

dépendant de paramètres réunis dans un vecteur $y = (y_1, \ldots, y_d) \in Y \subset \mathbb{R}^d$. Un exemple simple déjà évoqué dans l'introduction, et qui sera approfondi dans le chapitre 5, est celui de l'équation de diffusion stationnaire

$$-\mathrm{div}(a\nabla u) = f, \quad \text{dans} \quad D \subset \mathbb{R}^m, \quad u_{|\partial D} = 0, \tag{4.35}$$

où $f \in L^2(D)$ et

$$a = a(y) = \overline{a} + \sum_{j=1}^{d} y_j \psi_j, \tag{4.36}$$

avec $\overline{a}$ et $\{\psi_1, \ldots, \psi_d\}$ des fonctions données définies sur D. Il faut ici bien faire la distinction entre les variables paramétriques $y \in Y$ et la variable spatiale $x \in D$ à laquelle se rapporte les opérateurs de différentiation div et ∇ dans l'équation : pour chaque $y \in Y$ fixé, la fonction de diffusion $a(y)$ et la solution $u(y)$ sont des fonctions qui dépendent de la variable $x \in D$. On note parfois,

$$a(x, y) = a(y)(x) \quad \text{et} \quad u(x, y) = u(y)(x), \quad y \in Y,\ x \in D, \tag{4.37}$$

mais dans notre étude on fera souvent abstraction de la variable spatiale. En supposant que

$$0 < r \leq a(x, y) \leq R < \infty, \quad y \in Y,\ x \in D, \tag{4.38}$$

le théorème de Lax-Milgram nous indique que pour tout $y \in Y$ il existe une unique solution

$$u(y) \in V := H_0^1(D), \tag{4.39}$$

où $H_0^1(D)$ est le sous espace des fonctions de l'espace de Sobolev $H^1(D)$ dont la trace $u_{|D}$ est nulle.

Pour de telles équations, la classe

$$\mathcal{K} = \{u(y) \ : \ y \in Y\} \subset V, \tag{4.40}$$

est parfois appellée *variété des solutions* (solution manifold), puisqu'on peut la voir comme une variété contenue dans V à d paramètres $y_1, \ldots, y_d$. On suppose que c'est un compact

de V ce qui est le cas lorsque Y est un compact de $\mathbb{R}^d$ et que l'application $y \mapsto u(y)$ est continue. On cherche des espaces V_n qui approchent l'infimum dans la définition de l'épaisseur $d_n = d_n(\mathcal{K})_V$.

La méthode des bases réduites consiste à choisir astucieusement n éléments $\{u^0, \ldots, u^{n-1}\} \subset \mathcal{K}$, c'est à dire n solutions particulières $u^i = u(y^i)$ dans le cadre applicatif qu'on vient de décrire, et à définir

$$V_n = \text{vect}\{u^0, \ldots, u^{n-1}\}. \tag{4.41}$$

Comme on l'a expliqué, ces solutions peuvent en pratique être calculées avec une précision prescrites dans un espace d'élément fini V_h de dimension N de sorte que $V_n \subset V_h$. L'objectif est que l'espace V_n permette d'approcher l'ensemble des solutions $u(y) \in \mathcal{K}$ avec une précision comparable à celle de V_h et une réduction de dimension $n << N$ significative. La méthode des bases réduites est utile dans les situations où on doit calculer approximativement la solution $u(y)$ pour de nombreuses valeurs de y. On distingue typiquement deux phases des calculs :

1. Dans une phase *offline*, les solutions $u^0, \ldots, u^{n-1}$ sont calculées dans l'espace V_h ce qui demande de déterminer $\mathcal{O}(nN)$ inconnues. Elles sont mise en mémoire une fois pour toute et cela détermine l'espace V_n.
2. Dans la phase *online*, pour toute valeur de y demandée, on utilise une méthode numérique (typiquement la méthode de Galerkin) pour calculer une approximation $u_n(y) \in V_n$ de $u(y)$, ce qui demande de déterminer $\mathcal{O}(n)$ inconnues.

La phase offline est donc coûteuse (chaque calcul de solution est fait dans un espace de dimension $N >> 1$) mais effectuée une seule fois, alors que chaque calcul de solution effectué dans la phase online est peu coûteux.

Expliquons à présent la méthode qu'on va utiliser pour le choix de $\{u^0, \ldots, u^{n-1}\}$. L'idée de base est que si on a déjà construit $V_k := \text{vect}\{u^0, \ldots, u^{k-1}\}$, il est inutile de prendre pour u^k un élément qui est déjà dans V_k. A contrario, on a envie de choisir celui qui s'en éloigne le plus, ce qui conduit à définir par récurrence

$$u^k := \text{argmax}\{\|u - P_{V_k} u\| \ : \ u \in \mathcal{K}\}, \tag{4.42}$$

ou de manière équivalente $u^k = u(y^k)$ où

$$y^k := \text{argmax}\{\|u(y) - P_{V_k} u(y)\| \ : \ y \in Y\} \tag{4.43}$$

En posant $V_0 = \{0\}$ on peut initialiser la procédure en choisissant u^1 maximisant $\|u\|$ sur $\mathcal{K}$. Cet algorithme greedy est une version idéalisée de ce qu'on va réellement faire : on ne peut pas se permettre de calculer exactement l'erreur de projection $\|u(y) - P_{V_k} u(y)\|$ pour tout $y \in Y$, car il faudrait pour cela calculer toutes les solutions $u(y)$, au moins approximativement dans l'espace V_h, ce qu'on cherche précisément à éviter ! En pratique, il est possible de calculer à un coût bien moindre pour chaque $y \in Y$ une quantité $E_k(y)$ qui est équivalente à l'erreur de projection au sens où il existe $0 < c_1 < c_2$ telles que

$$c_1 E_k(y) \leq \|u(y) - P_{V_k} u(y)\| \leq c_2 E_k(y), \quad y \in Y, \ k \geq 1. \tag{4.44}$$

Une telle quantité peut par exemple être définie à partir de la solution numérique $y \mapsto u_k(y) \in V_k$ dont on a vu qu'elle était peu coûteuse, en calculant son résidu

$$E_k(y) := \|\mathcal{P}(u_k(y), y)\|_*, \tag{4.45}$$

où $\|\cdot\|_*$ est une norme bien choisie. Dans le cas de l'équation de diffusion elliptique (4.35), on pourra vérifier (**exercice difficile**) que

$$E_k(y) := \|f + \mathrm{div}(a\nabla u_k(y))\|_{H^{-1}}, \tag{4.46}$$

convient, où $H^{-1}(D)$ est le dual de $H^1_0(D)$ et

$$\|v\|_{H^{-1}} = \max\Big\{ \int_D vw \ : \ \|w\|_{H^1_0} = \|\nabla w\|_{L^2} = 1\Big\}. \tag{4.47}$$

Disposant d'une telle quantité, on modifie l'algorithme greedy en définissant

$$y^k := \mathrm{argmax}\{E_k(y) \ : \ y \in Y\} \tag{4.48}$$

L'équivalence (4.44) nous assure que

$$\|u(y^k) - P_{V_k}u(y^k)\| \geq c_1 E_k(y^k) \geq c_1 E_k(y) \geq \frac{c_1}{c_2}\|u(y) - P_{V_k}u(y)\|, \quad y \in Y. \tag{4.49}$$

En posant $\gamma := \frac{c_1}{c_2} \in]0,1[$, cela revient donc à modifier chaque étape de l'algorithme greedy de la manière suivante : on choisi $u^k \in \mathcal{K}$ tel que

$$\|u^k - P_{V_k}u^k\| \geq \gamma\|u - P_{V_k}u\|, \quad u \in \mathcal{K}. \tag{4.50}$$

Nous allons étudier en toute généralité cette version plus réaliste, qui est parfois appelée algorithme greedy *faible*, de paramètre γ.

Remarque 4.5 *Une autre modification de l'algorithme dont nous n'analyserons pas l'effet ici concerne le calcul du maximum de $y \mapsto E_k(y)$. Cette fonction n'étant pas concave et pouvant posséder de multiples maxima locaux, les algorithmes classiques d'optimisation continue peuvent être mis en défaut. Une approche parfois suivie est de restreindre la maximisation à un sous ensemble $\widetilde{Y} \subset Y$ fini qui a la forme d'un réseau de points suffisamment fin. Cela revient à remplacer $\mathcal{K}$ dans (4.50) par un sous-ensemble $\widetilde{\mathcal{K}}$ qui est une discrétisation suffisament fine de $\mathcal{K}$.*

Si $(V_n)_{n\geq 1}$ est la suite d'espaces engendrés par l'algorithme greedy, on mesurera la performance de ces espaces par

$$\sigma_n = \sigma_n(\mathcal{K})_V = \max_{u\in\mathcal{K}} \|u - P_{V_n}u\|. \tag{4.51}$$

Notons que tout comme $(d_n)_{n\geq 0}$ la suite $(\sigma_n)_{n\geq 0}$ est décroissante et qu'on a

$$\sigma_0 = d_0 = \max_{u\in\mathcal{K}} \|u\|. \tag{4.52}$$

La propriété de sélection de l'algorithme greedy faible nous indique que

$$\gamma\sigma_n \leq \|u^n - P_{V_n}u^n\| \leq \sigma_n. \tag{4.53}$$

Par ailleurs, il est facile de montrer que la suite σ_n doit tendre vers 0 : si ce n'était pas le cas on montre que la suite $(u^n)_{n\geq 0}$ ne contient pas de sous-suite convergente (**exercice**) ce qui contredirait la compacité de $\mathcal{K}$.

On a évidemment $\sigma_n \geq d_n$, et nous allons chercher à comprendre si σ_n est néammoins du même ordre que d_n, auquel cas les espaces V_n pourront être qualifiés de presque-optimaux.

Un cas simple donnant l'intuition du bien fondé de cette analyse est celui où $\mathcal{K}$ est un ellipsoide associé à une base orthonormée $(\varphi_j)_{j\geq 0}$ et une suite de longueurs de demi-axes $r_0 > r_1 > \cdots \to 0$, c'est à dire

$$\mathcal{K} = \Big\{ u = \sum_{j\geq 0} u_j \varphi_j \ : \ \sum_{j\geq 0} r_j^{-2} |u_j|^2 \leq 1 \Big\}. \tag{4.54}$$

Dans ce cas, lorsqu'on applique l'algorithme greedy idéalisé (4.42), il est facile de montrer (**exercice**) que la suite choisie sera

$$u^j = s_j \varphi_j, \quad j \geq 0 \tag{4.55}$$

et que les espaces $V_n = \mathrm{vect}\{\varphi_0, \ldots, \varphi_{n-1}\}$ sont optimaux au sens où $\sigma_n = d_n = r_n$. Cependant cette propriété d'optimalité n'est plus vérifiée lorsque $\mathcal{K}$ a une géométrie plus générale (on peut aisément s'en rendre compte à partir de quelques exemples simples), et une analyse beaucoup plus fine est nécessaire.

4.3 Analyse et résultats de convergence

Avant d'entamer l'analyse de la méthode, donnons un avant goût des résultats de convergence qu'on va obtenir. Idéalement on aimerait prouver qu'il existe une constante $C > 1$ telle que

$$\sigma_n \leq C d_n, \quad n \geq 1, \tag{4.56}$$

pour tout compact $\mathcal{K} \subset V$. Le premier résultat de comparaison entre σ_n et d_n obtenu par Buffa-Maday-Patera-Turinici pour l'algorithme greedy idéalisé, est très loin de cet objectif puisqu'il affirme que

$$\sigma_n \leq n 2^n d_n, \quad n \geq 1. \tag{4.57}$$

La présence d'un facteur exponentiel est en fait inévitable : on peut montrer que pour tout $\varepsilon > 0$ et pour tout $n \geq 1$, il existe un compact $\mathcal{K}$ tel que

$$\sigma_n \geq (1-\varepsilon) 2^n d_n. \tag{4.58}$$

Ce résultat anihile tout espoir d'un résultat général de type (4.56).

La situation devient cependant beaucoup plus favorable si on cherche à formuler la comparaison en termes de vitesse de convergence. Le résultat suivant (dont on peut s'assurer qu'il n'est pas en contradiction avec le résultat négatif précèdent), obtenu par Binev-Cohen-Dahmen-DeVore-Petrova-Wojtacztzyck en 2011, affirme que les espaces V_n de bases réduites engendrés par l'algorithme greedy faible sont optimaux au sens de la vitesse d'approximation lorsque $n \to \infty$.

Théorème 4.2 *Si pour une constante C, les épaisseurs de Kolmogorov de $\mathcal{K}$ vérifie $d_0 \leq C$ et*

$$d_n \leq Cn^{-s}, \quad n \geq 1, \tag{4.59}$$

alors la performance de l'algorithme greedy faible vérifie

$$\sigma_n \leq \tilde{C}n^{-s}, \quad n \geq 1, \tag{4.60}$$

où $\tilde{C}$ depend de (C, s, γ).

Afin d'établir ce type de résultat, on va reformuler l'étude de l'algorithme greedy faible sous une forme matricielle. Si $(u^j)_{j\geq 0}$ est la suite des vecteurs sélectionnés par l'algorithme, on introduit la suite $(v^j)_{j\geq 0}$ obtenue par le procédé de Gramm-Schmidt, c'est à dire $v^0 := \|u^0\|^{-1}u^0$ et

$$v^j = \frac{1}{\|u^j - P_{V_j}u^j\|}(u^j - P_{V_j}u^j), \quad j \geq 1. \tag{4.61}$$

Par construction $V_n = \text{vect}\{v^0, \ldots, v^{n-1}\}$ et pour chaque $i \geq 0$ on peut donc écrire

$$u^i = \sum_{j=0}^{i} a_{i,j}v^j, \tag{4.62}$$

où

$$a_{i,j} = \langle u^i, v^j \rangle. \tag{4.63}$$

On introduit la matrice triangulaire bi-infinie qui décrit ce changement de base

$$\mathbf{A} = (a_{i,j})_{i,j\geq 0} = \begin{pmatrix} a_{0,0} & 0 & 0 & 0 & \ldots & \ldots & \ldots & \ldots \\ a_{1,0} & a_{1,1} & 0 & 0 & \ldots & \ldots & \ldots & \ldots \\ \vdots & \ddots & \ddots & \ddots & \ddots & & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & & & \\ a_{n,0} & a_{n,1} & \ldots & \ldots & a_{n,n} & 0 & 0 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \end{pmatrix} \tag{4.64}$$

Cette matrice possède deux propriétés fondamentales pour l'analyse qui va suivre :

1. Pour tout n on a $a_{n,n}v^n = u^n - P_{V_n}u^n$ et par conséquence $|a_{n,n}| = \|u^n - P_{V_n}u^n\|$. La propriété de sélection (4.53) de l'algorithme greedy implique ainsi la propriété

   (P1) $$\gamma\sigma_n \leq |a_{n,n}| \leq \sigma_n, \quad n \geq 0.$$

   L'étude de la décroissance de σ_n se ramène ainsi à celle des éléments diagonaux de $\mathbf{A}$.

2. Pour $m \geq n$, on peut écrire $\|u_m - P_{V_n} u_m\|^2 = \sum_{j=n}^{m} |\langle u^m, v^j \rangle|^2 = \sum_{j=n}^{m} a_{m,j}^2$. Et en particulier

$$\text{(P2)} \qquad \sum_{j=n}^{m} a_{m,j}^2 \leq \sigma_n^2, \qquad m \geq n.$$

Comme $\sigma_n^2 \leq \gamma^{-2} |a_{n,n}|^2$, cette propriété peut s'interprêter de la manière suivante : si on considère une sous-matrice triangulaire extraite de $\mathbf{A}$ en conservant les lignes et colonnes pour $n \leq i, j \leq m$, la norme ℓ^2 de chaque ligne de cette matrice est dominée par celle de la première ligne, au facteur γ^{-2} près.

Ces deux propriétés caractérisent en fait toutes les matrices qui peuvent être issues de l'algorithme greedy faible : pour toute matrice triangulaire $\mathbf{A}$ satisfaisant (P1) et (P2) pour une suite $(\sigma_n)_{n \geq 0}$ qui tend vers 0, il existe un ensemble compact $\mathcal{K}$ dans un espace de Hilbert V tel qu'une réalisation de l'algorithme greedy faible produit la matrice $\mathbf{A}$. Il suffit de se placer dans $V = \ell^2(\mathbb{N})$ et de prendre pour $\mathcal{K}$ l'ensemble des lignes

$$\mathbf{a}_i = (a_{i,0}, \dots, a_{i,i}, 0, 0, \dots), \quad i \geq 0, \tag{4.65}$$

de la matrice $\mathbf{A}$. L'algorithme greedy faible peut alors sélectionner dans l'ordre $u^i = \mathbf{a}_i$ dont l'orthogonalisation par Gramm-Schmidt donne alors la base canonique $v^j = \mathbf{e}_j = (0, \dots, 0, 1, 0, \dots)$.

Notons que la base orthonormale $(v^j)_{j \geq 0}$ permet de définir une isométrie $L : V \to \ell^2(\mathbb{N})$ en écrivant

$$Lu = \mathbf{c} = (c_j)_{j \geq 0} \quad \text{pour} \quad u = \sum_{j \geq 0} c_j v^j, \tag{4.66}$$

et qu'on a en particulier $Lu^i = \mathbf{a}_i$.

On veut maintenant comparer la suite $(\sigma_n)_{n \geq 0}$, ou de manière équivalente les éléments diagonaux $|a_{n,n}|$, avec la suite $(d_n)_{n \geq 0}$ des épaisseurs de Kolmogorov. Pour cela, on note que pour tout $m \geq 0$ fixé, si $\overline{V}_m$ est un espace optimal au sens où

$$\max_{u \in \mathcal{K}} \|u - P_{\overline{V}_m} u\| = d_m, \tag{4.67}$$

on a en particulier,

$$\|u^i - P_{\overline{V}_m} u^i\| \leq d_m, \quad i \geq 0. \tag{4.68}$$

En appliquant l'isométrie L, on peut définir un espace de dimension m

$$W_m = L(\overline{V}_m) \subset \ell^2(\mathbb{N}), \tag{4.69}$$

et on a ainsi de manière équivalente

$$\|\mathbf{a}_i - P_{W_m} \mathbf{a}_i\|_{\ell^2} \leq d_m, \quad i \geq 0. \tag{4.70}$$

Tous les vecteurs lignes de la matrice $\mathbf{A}$ sont donc à distance au plus d_m dans ℓ^2 d'un espace de dimension m. Ceci reste évidemment vrai si on considére des versions restreintes de ces vecteurs, par exemple les vecteurs

$$\tilde{\mathbf{a}}_i = (a_{i,k+1}, a_{i,k+2}, \dots, a_{i,k+K}) \in \mathbb{R}^K, \quad i \geq 0, \tag{4.71}$$

pour k et K fixés, puisqu'il suffit de considérer l'espace W de dimension au plus m obtenu en prenant les éléments de W_m et en leur appliquant la même restriction qui diminue la norme.

On a vu qu'on ne pouvait pas toujours garantir l'existence d'un espace optimal $\overline{V}_m$ qui atteint l'infimum d_m dans la définition de l'épaisseur de Kolmogorov, mais on peut toujours en trouver un qui atteint la précision $\tilde{d}_m$ pour n'importe quel $\tilde{d}_m > d_m$. En d'autre termes, on a obtenu la troisième propriété fondamentale suivante :

(P3) *Pour tout $m \geq 0$ et $\tilde{d}_m > d_m$, et pour toute sous-matrice $\mathbf{B}$ de $\mathbf{A}$ de largeur K, il existe un sous-espace W de dimension au plus m tel que chaque ligne de $\mathbf{B}$ est à distance au plus $\tilde{d}_m$ de W en norme ℓ^2.*

Afin d'exploiter cette propriété, on va avoir recours à un résultat d'analyse matricielle dû a DeVore-Petrova-Wojtacztzyck, dont nous différons la preuve pour en donner d'abord les conséquences.

Lemme 4.1 *Soit $\mathbf{B}$ une matrice $K \times K$ et soit $W \subset \mathbb{R}^K$ un sous-espace de dimension $m \leq K$, on a alors*

$$\det(\mathbf{B})^2 \leq \Big(\frac{1}{m}\sum_{i=1}^{K} \|P_W \mathbf{b}_i\|_{\ell^2}^2\Big)^m \Big(\frac{1}{K-m}\sum_{i=1}^{K} \|\mathbf{b}_i - P_W \mathbf{b}_i\|_{\ell^2}^2\Big)^{K-m}, \tag{4.72}$$

où P_W est le projecteur orthogonal sur W et $\mathbf{b}_1, \ldots, \mathbf{b}_K \in \mathbb{R}^K$ sont les lignes de $\mathbf{B}$.

Nous allons ce lemme à la matrice $\mathbf{B}$ définie par restriction de $\mathbf{A}$ aux indices $i, j = N+1, \ldots, N+K$, qui est une matrice $K \times K$ triangulaire pour laquelle on a, par (P1),

$$\det(\mathbf{B})^2 = \prod_{i=1}^{K} a_{N+i,N+i}^2 \geq \gamma^{2K} \prod_{i=1}^{K} \sigma_{N+i}^2. \tag{4.73}$$

Par (P3), on sait qu'il existe un espace W de dimension m tel que

$$\|\mathbf{b}_i - P_W \mathbf{b}_i\|_{\ell^2} \leq \tilde{d}_m, \quad i = 1, \ldots, K. \tag{4.74}$$

Par (P2), on peut aussi écrire

$$\|P_W \mathbf{b}_i\|_{\ell^2} \leq \|\mathbf{b}_i\|_{\ell^2} \leq \sigma_{N+1}. \tag{4.75}$$

En combinant ces estimations avec le Lemme et en faisant tendre $\tilde{d}_m$ vers d_m, on obtient ainsi et

$$\gamma^{2K} \prod_{i=1}^{K} \sigma_{N+i}^2 \leq \Big(\frac{K\sigma_{N+1}^2}{m}\Big)^m \Big(\frac{K d_m^2}{K-m}\Big)^{K-m}. \tag{4.76}$$

En posant $x = \frac{m}{K} \in]0, 1]$, et en utilisant la décroissance de σ_n, on obtient

$$\gamma^{2K} \sigma_{N+K}^{2K} \leq \Big(x^{-x}(1-x)^{x-1}\Big)^K (\sigma_{N+1})^{2m} (d_m)^{2(K-m)}. \tag{4.77}$$

Une étude de fonction permet de vérifier que $x^{-x}(1-x)^{x-1} \leq 2$ pour $x \in]0,1]$, et en élevant à la puissance $\frac{1}{2K}$ on obtient finalement l'estimation

$$\sigma_{N+K} \leq \beta \left(\sigma_{N+1}\right)^{\frac{m}{K}} \left(d_m\right)^{1-\frac{m}{K}}, \qquad \beta = \frac{\sqrt{2}}{\gamma}. \tag{4.78}$$

On peut maintenant obtenir des résultats de comparaison entre les vitesses de décroissances de d_n et σ_n en appliquant cette estimation à des valeurs particulières de N et K.

Prenons d'abord $N = 0$, $m = n$ et $K = 2n$. En remarquant que $\sigma_1 \leq \sigma_0 = d_0$, cela donne

$$\sigma_{2n} \leq \beta \sqrt{d_0 d_n}. \tag{4.79}$$

Cette inégalité simple nous permet une première comparaison dans le cas de vitesses exponentielles de la forme

$$d_n \leq C \exp(-cn^s), \quad n \geq 0. \tag{4.80}$$

Dans ce cas on obtient

$$\sigma_{2n} \leq \beta C \exp\Big(-\frac{c}{2} n^s\Big), \tag{4.81}$$

ce qui équivaut à l'estimation

$$\sigma_k \leq \beta C \exp\Big(-\frac{c}{2^{s+1}} k^s\Big), \tag{4.82}$$

pour toute les valeurs paires de k. En utilisant le fait que $\sigma_{2n+1} \leq \sigma_{2n}$, on obtient facilement une estimation du même type pour les valeurs impaire quitte à modifier légèrement les constantes. On a ainsi obtenu le résultat suivant.

Théorème 4.3 *Si pour une constante C, les épaisseurs de Kolmogorov de $\mathcal{K}$ vérifient et*

$$d_n \leq C \exp(-cn^s), \quad n \geq 0. \tag{4.83}$$

alors la performance de l'algorithme greedy faible vérifie

$$\sigma_n \leq \tilde{C} \exp(-\tilde{c} n^s), \quad n \geq 0. \tag{4.84}$$

où $(\tilde{C}, \tilde{c})$ dépendent de (C, c, s, γ).

Pour établir le Theorème 4.2 qui concerne les vitesses de convergence polynomiales, on applique (4.78) aux valeurs $N = K = 2n$ et $m = n$ ce qui donne

$$\sigma_{4n} \leq \beta \sqrt{\sigma_{2n} d_n}. \tag{4.85}$$

On considère d'abord les valeurs $n = 2^j$, et l'hypothèse du théorème entraine en particulier que

$$d_{2^j} \leq C 2^{-js}, \quad j \geq 0. \tag{4.86}$$

On va montrer que

$$\sigma_{2^j} \leq \tilde{C} 2^{-sj}, \quad j \geq 0, \tag{4.87}$$

par récurrence sur j pour une constante $\tilde{C}$ bien choisie. On a $\sigma_2 \leq \sigma_1 \leq \sigma_0 \leq C$. Et pour $j \geq 1$, on peut écrire

$$\sigma_{2^{j+1}} \leq \beta\sqrt{\sigma_{2^j} d_{2^{j-1}}} \leq \beta\sqrt{\tilde{C}C}\sqrt{2^{-s(2j-1)}} \leq \beta\sqrt{2^{3s}\tilde{C}C}\; 2^{-s(j+1)}, \tag{4.88}$$

La récurrence sera donc vérifiée si $\tilde{C} = \beta\sqrt{2^{3s}\tilde{C}C}$, soit

$$\tilde{C} = \beta^2 2^{3s} C. \tag{4.89}$$

Comme $\beta > 1$, on a $C \leq \tilde{C}2^{-3s}$ ce qui montre a-postériori que la propriété recherchée est aussi vérifiée pour les valeurs d'initialisation $j = 0, 1$. Enfin, pour les valeurs $2^j < n < 2^{j+1}$, on obtient

$$\sigma_n \leq \sigma_{2^j} \leq \tilde{C}2^{-sj} \leq 2^s \tilde{C} 2^{-s(j+1)} \leq 2^s \tilde{C} n^{-s}, \tag{4.90}$$

ce qui prouve le Théorème 4.2 avec la constante modifiée $\tilde{C} = \beta^2 2^{4s} C$.

Il nous reste à prouver le Lemme 4.1. Pour cela, on considère une base orthonormée $\mathbf{w}_1, \ldots, \mathbf{w}_m$ de l'espace W et on la complete en une base orthonormée $\{\mathbf{w}_1 \ldots, \mathbf{w}_K\}$ de $\mathbb{R}^K$. On dénote par $\mathbf{M}$ la matrice de changement de base dont la colonne j est donnée par les $\mathbf{w}_j^T$ et la matrice

$$\mathbf{T} := \mathbf{BM} = (\langle \mathbf{b}_i, \mathbf{w}_j \rangle)_{i,j=1,\ldots,K}. \tag{4.91}$$

Puisque $\mathbf{M}$ est unitaire, on a

$$\det(\mathbf{B})^2 = \det(\mathbf{T})^2. \tag{4.92}$$

L'inégalité de Hadamard nous indique que le déterminant d'une matrice est inférieur en valeur absolue au produit des normes ℓ^2 de ses colonnes, avec égalité lorsque celles-ci sont orthogonales (**exercice**). Ceci entraine donc

$$\det(\mathbf{B})^2 \leq \prod_{j=1}^{K} N_j, \quad N_j = \sum_{i=1}^{K} |\langle \mathbf{b}_i, \mathbf{w}_j \rangle|^2. \tag{4.93}$$

On note que

$$\sum_{j=1}^{m} N_j = \sum_{i=1}^{K}\sum_{j=1}^{m} |\langle \mathbf{b}_i, \mathbf{w}_j \rangle|^2 = \sum_{i=1}^{K} \|P_W \mathbf{b}_i\|^2, \tag{4.94}$$

et

$$\sum_{j=m+1}^{K} N_j = \sum_{i=1}^{K}\sum_{j=m+1}^{K} |\langle \mathbf{b}_i, \mathbf{w}_j \rangle|^2 = \sum_{i=1}^{K} \|\mathbf{b}_i - P_W \mathbf{b}_i\|^2. \tag{4.95}$$

On sépare le produit en deux blocs $j = 1, \ldots, m$ et $j = m+1, \ldots, K$ et on applique l'inégalité entre la moyenne géométrique et arithmétique qui donne

$$\prod_{j=1}^{K} N_j \leq \Big(\frac{1}{m}\sum_{j=1}^{m} N_j\Big)^m \Big(\frac{1}{K-m}\sum_{j=m+1}^{K} N_j\Big)^{K-m}. \tag{4.96}$$

En combinant avec les deux égalités ci-dessus, on a ainsi établi le lemme.

Remarque 4.6 *Il est possible de généraliser l'algorithme greedy faible au cas où V est un espace de Banach, mais on ne sait pas prouver dans ce cadre plus général que les espaces V_n ont la même vitesse d'approximation que les espaces optimaux des épaisseurs de Kolmogorov.*

5 Approximation en dimension infinie

Dans ce dernier chapitre nous allons nous intéresser à un problème d'approximation pour une fonction à très grand nombre (potentiellement infini et dénombrable) de variables, issue d'une EDP elliptique paramétrique. Cet exemple va permettre d'illustrer les mécanismes d'anisotropie qui permettent de contourner la malédiction de la grande dimension, ainsi que l'utilisation des méthodes parcimonieuses dans ce cadre. Il nous conduira aussi à des résultats sur les épaisseurs de Kolmogorov des variétés solutions pour ce type d'EDP.

5.1 Approximation polynomiale d'une EDP paramétrique

Nous considérons à nouveau le problème aux limites donné par l'EDP elliptique

$$-\mathrm{div}(a\nabla u) = f, \quad \text{dans} \quad D \subset \mathbb{R}^m, \quad u_{|\partial D} = 0, \tag{5.1}$$

posée sur un domaine $D \subset \mathbb{R}^m$, avec une fonction de diffusion a qui dépend d'un nombre arbitrairement grand de variables paramétriques, sous la forme affine

$$a = a(y) = \overline{a} + \sum_{j\geq 1} y_j \psi_j, \tag{5.2}$$

avec $\overline{a}$ et $(\psi_j)_{j\geq 1}$ des fonctions fixées dans $L^\infty(D)$. On autorise ainsi un nombre infini de variables, le cas fini à d variables correspondant à imposer que $\psi_j = 0$ quand $j > d$.

On peut penser par exemple à $\sum_{j\geq 1} y_i\psi_j$ comme à un développement de $a-\overline{a}$ en séries de Fourier, ou à un développement de Karhunen-Loeve dans le cas où a est un champs aléatoire et $\overline{a} = \mathbb{E}(a)$, ou à un développement suivant les indicatrices $\psi_j = \chi_{D_j}$ lorsque a est constant par morceaux sur une partition $\{D_1, \dots, D_d\}$ de D.

Pour simplifier notre étude on supposera que les variables y_j varient chacune dans l'intervalle $[-1, 1]$, c'est à dire que

$$y = (y_j)_{j\geq 1} \in Y = [-1, 1]^{\mathbb{N}}. \tag{5.3}$$

Notons qu'on peut toujours se ramener à cette situation lorsque les y_j varient chacun dans des intervalles I_j plus généraux (c'est à dire $Y = I_1 \times I_2 \times \dots$), en faisant un changement de variable affine sur chaque y_j ce qui revient à modifier les fonctions $\overline{a}$ et ψ_j. Il s'agit tout de même d'une hypothèse simplificatrice (qui sera utile dans l'analyse qu'on va développer plus loin) car on pourrait considérer des domaines paramétriques Y plus généraux qui n'ont pas une forme rectangulaire.

Nous ferons ici l'hypothèse que la série (5.2) est convergente dans $L^\infty(D)$, uniformément en la variable y, c'est à dire

$$\lim_{J\to\infty} \sup_{y\in Y} \|a(y) - a_J(y)\|_{L^\infty(D)} = 0, \tag{5.4}$$

où $a_J(y) := \overline{a} + \sum_{j=1}^J y_j\psi_j$ est la série tronquée.

Rappelons quelques résultats fondamentaux sur le problème aux limites elliptique (5.1) : il est judicieux d'introduire le sous espace hilbertien

$$V = H_0^1(D) := \{v \in H^1(D) \ : \ v_{|\partial D} = 0\}, \tag{5.5}$$

la théorie des espaces de Sobolev permettant de donner un sens à la trace sur ∂D. On définit sur cet espace la norme

$$\|v\|_V := \|\nabla v\|_{L^2}, \tag{5.6}$$

qui est équivalente sur V à la norme H^1 en vertu de l'inégalité de Poincaré

$$\|v\|_{L^2} \leq C_P \|\nabla v\|_{L^2}, \quad v \in V. \tag{5.7}$$

On passe à la *formulation variationelle* en multipliant l'équation (5.1) par n'importe quelle fonction $v \in V$ et en intégrant sur D ce qui par intégration par partie (formule de Green) et annulation du terme de bord donne

$$\int_D a\nabla u \nabla v = -\int_D \mathrm{div}(a\nabla u)v = \int_D fv, \quad v \in V. \tag{5.8}$$

Cette formulation garde ainsi un sens lorsque u a seulement des dérivées d'ordre 1 dans L^2, et on a ainsi mis le problème sous la forme : trouver $u \in V$ tel que

$$A(u, v) = L(v), \qquad v \in V, \tag{5.9}$$

où $A(u, v) = \int_D a\nabla u \nabla v$ et $L(v) = \int_D fv$. Sous l'hypothèse d'ellipticité $0 < r \leq a \leq R < \infty$, la forme bilinéaire A vérifie les propriétés de continuité et de coercivité sur V,

$$|A(u, v)| \leq R\|u\|_V \|v\|_V \quad \text{et} \quad A(u, u) \geq r\|u\|_V^2, \quad u, v \in V, \tag{5.10}$$

et la forme linéaire L est continue sur V,

$$|L(v)| \leq \|f\|_{L^2}\|v\|_{L^2} \leq C_L\|v\|_V, \quad C_L = C_P\|f\|_{L^2}. \tag{5.11}$$

Sous ces trois hypothèses, le théorème de Lax-Milgram assure l'existence et l'unicité d'une solution $u \in V$. En prenant $v = u$ dans (5.9) on obtient en particulier

$$r\|u\|_V^2 \leq A(u, u) = L(u) \leq C_L\|u\|_V, \tag{5.12}$$

c'est à dire l'estimation a-priori

$$\|u\|_V \leq M, \qquad M := \frac{C_L}{r} = \frac{C_P\|f\|_{L^2}}{r}. \tag{5.13}$$

Si on se place à présent dans le cadre où $a = a(y)$ a la forme affine (5.2), la solution $u(y) \in V$ est solution de

$$A_y(u(y), v) = L(v), \qquad v \in V, \tag{5.14}$$

où

$$A_y(u, v) := \int_D a(y)\nabla u \nabla v = \int_D \overline{a}\nabla u \nabla v + \sum_{j\geq 1} y_j \int_D \psi_j \nabla u \nabla v, \tag{5.15}$$

est donc une forme bilinéaire qui varie avec $y \in Y$. On fait ici l'hypothèse d'ellipticité uniforme

$$0 < r \leq \overline{a}(x) + \sum_{j \geq 1} y_j \psi_j(x) \leq R < \infty, \quad x \in D, y \in Y, \tag{5.16}$$

qui nous assure par Lax-Milgram que, pour tout $y \in Y$, il existe une unique solution $u(y) \in V$ de (5.14). Ceci nous permet de définir *l'application solution* (parameter to solution map)

$$u : y \in Y \mapsto u(y) \in V. \tag{5.17}$$

Notons que l'hypothèse d'ellipticité uniforme entraîne la borne uniforme

$$\|u(y)\|_V \leq M, \quad y \in Y, \quad M := \frac{C_P \|f\|_{L^2}}{r}, \tag{5.18}$$

c'est à dire que $u \in L^\infty(Y, V)$, l'espace des fonctions bornées sur Y à valeur dans V.

C'est cette application que nous allons chercher à approcher numériquement, par des applications dont la dépendance en y est simple. Plus précisément, on va utiliser des polynômes en les variables y_j, c'est à dire des applications $u_n : Y \to V$ de la forme

$$u_n(y) = \sum_{\nu \in \Lambda_n} u_\nu y^\nu, \tag{5.19}$$

où

$$y^\nu = \prod_{j \geq 1} y_j^{\nu_j}, \quad \nu = (\nu_j)_{j \geq 1}, \tag{5.20}$$

et où Λ_n est un ensemble de multi-indices de cardinal $n = \#(\Lambda_n) < \infty$.

Comme pour la méthode des bases réduites, notre motivation est que l'évaluation précise de $u(y)$ par une méthode d'approximation classique (éléments finis, différence finies) devient coûteuse si on doit l'effectuer pour de nombreuses valeurs de y. La construction de telles approximation polynomiales soulève des difficultés aussi bien théoriques que numériques.

La première difficulté vient du fait que l'application $y \mapsto u(y)$ n'est pas à valeur dans $\mathbb{R}$ mais dans un espace de Hilbert $V = H_0^1(D)$ qui est de dimension infinie. De ce fait, les "coefficients polynomiaux" u_ν ne sont pas des nombre réels mais des éléments de V. En pratique cette difficulté est contournée de manière similaire à ce qu'on a fait pour la méthode des bases réduites : on se place dans un espace d'éléments finis V_h de dimension $N >> 1$ suffisament grande (c'est à dire $h << 1$ suffisament petit) pour que l'approximation de Galerkin $u_h(y) \in V_h$ solution de

$$A_y(u_h(y), v_h) = L(v_h), \quad v_h \in V_h, \tag{5.21}$$

soit de précision acceptable. On rappelle en particulier que le lemme de Cea assure pour chaque y l'estimation optimale

$$\|u(y) - u_h(y)\|_V \leq C \min_{v_h \in V_h} \|u(y) - v_h\|_V, \quad C = (R/r)^{1/2}. \tag{5.22}$$

On peut ainsi se ramener alors au problème d'approcher l'application

$$y \in Y \mapsto u_h(y) \in V_h, \tag{5.23}$$

par un polynôme

$$u_{n,h}(y) = \sum_{\nu \in \Lambda_n} u_{\nu,h} y^\nu. \tag{5.24}$$

Il faut noter l'analogie avec la méthodes des bases réduites : les coefficients $\{u_{\nu,h}\}$ sont n fonctions de l'espace V_h que l'on peut calculer et mettre en mémoire avec un budget nN dans une phase offline. Dans la phase online, on calcule pour $y \in Y$ une solution approchée

$$y \mapsto u_{n,h}(y) \in V_n = \text{vect}\{u_{\nu,h} \ : \ \nu \in \Lambda_n\}. \tag{5.25}$$

Celle-ci consiste simplement en une recomposition des $u_{\nu,h}$ suivant la formule (5.24), alors que dans la méthode des bases réduite on applique la méthode de Galerkin dans V_n qui exige la résolution d'un système $n \times n$. Dans la suite de notre discussion, pour simplifier les notations, nous ferons abstraction de cette étape de discrétisation dans V_h pour concentrer nos efforts sur l'étude de l'approximation polynomiale u_n de l'application u, les techniques et l'analyse étant exactement les mêmes pour celle de l'approximation $u_{n,h}$ de l'application u_h.

La deuxième difficulté vient du fait que l'application $y \mapsto u(y)$ dépend d'un nombre infini de variables, et on est donc confronté de plein fouet à la malédiction de la dimensionalité. On voit par exemple qu'on ne peut envisager d'utiliser des polynômes de degré fixé en toutes les variables, par exemple de la forme

$$\sum_{\max_{j\geq 1} \nu_j \leq k} u_\nu y^\nu, \tag{5.26}$$

où de degré total fixé

$$\sum_{|\nu| = \sum_{j\geq 1} \nu_j \leq k} u_\nu y^\nu, \tag{5.27}$$

car dans les deux cas de tels polynômes comporteraient un nombre infini de termes. Notons que ces espaces polynomiaux sont aussi proscrits lorsque qu'il y a qu'un nombre fini mais très grand de variables $d >> 1$, puisque le nombre de termes est alors $(k+1)^d$ dans le premier cas et $\binom{k+d}{k}$ dans le deuxième cas qui dépendent tout deux exponentiellement de d. Afin d'obtenir des espaces plus "économiques", il sera crucial d'introduire de l'*anisotropie*, c'est à dire de ne pas avoir les mêmes degrés maximaux pour toutes les variables. En premier lieu, on choisira toujours les indices ν dans le sous ensemble des suites d'entiers positifs ou nuls à support fini, c'est à dire

$$\mathcal{F} = \ell^0(\mathbb{N}^*, \mathbb{N}) = \Big\{ \nu = (\nu_j)_{j\geq 1} \ : \ \nu_j \in \mathbb{N} \quad |\nu| = \sum_{j\geq 0} \nu_j < \infty \Big\}. \tag{5.28}$$

Avec une telle convention, on voit en particulier que les produits qui définissent y^ν sont toujours finis, puisque

$$y^\nu = \prod_{j \ : \ \nu_j \neq 0} y_j^{\nu_j}. \tag{5.29}$$

Si Λ_n est un sous ensemble fini de $\mathcal{F}$, on voit aussi qu'un polynôme u_n de la forme (5.19) ne dépend en fait que d'un nombre fini de variables y_j : en posant

$$J = \min\{j \ : \ \nu_l = 0, \ l > j, \ \nu \in \Lambda_n\}, \tag{5.30}$$

on voit que u_n ne dépend que de $(y_1, \dots, y_J)$ et est constante en les variables y_l pour $l > J$.

5.2 Un développement en série entière

Il existe de nombreuses manières de définir un polynôme d'approximation en une ou plusieurs variables : projection, interpolation... Dans notre analyse, on va construire u_n comme un développement de Taylor autour de $y = 0$, c'est à dire par troncature d'une série entière multivariée de la forme

$$\sum_{\nu \in \mathcal{F}} u_\nu y^\nu, \tag{5.31}$$

en conservant un nombre n de termes bien choisi, ce qui nous ramènera au problème de l'approximation à n-terme qu'on a déjà discuté en toute généralité. Dans un développement de Taylor multivarié, rappelons qu'on a

$$u_\nu := \frac{1}{\nu!} \partial^\nu u(0), \qquad \partial^\nu u := \frac{\partial^{|\nu|} u}{\partial^{\nu_1} y_1 \partial^{\nu_2} y_2 \dots}, \quad \nu! = \prod_{j \geq 1} \nu_j !, \tag{5.32}$$

avec la convention $0! = 1$. Il convient donc en premier lieu de donner un sens aux dérivées partielles de $y \mapsto u(y)$ à tous les ordres et à la convergence de la série entière ci-dessus vers $u(y)$.

Notre point de départ est un résultat de stabilité Lipschitz de la solution de l'EDP (5.1) vis à vis de la fonction de diffusion.

Lemme 5.1 *Si a_1 et a_2 sont deux fonctions de diffusion sur D telles que*

$$0 < r \leq a_i \leq R < \infty, \quad i = 1, 2 \tag{5.33}$$

et si u_1 et u_2 désignent les solutions de (5.1) pour $a = a_1$ et $a = a_2$, alors on a

$$\|u_1 - u_2\|_V \ \leq C \|a_1 - a_2\|_{L^\infty}, \tag{5.34}$$

où $C = \frac{M}{r}$ avec $M = \frac{C_P \|f\|_{L^2}}{r}$.

Preuve : on écrit pour tout $v \in V$

$$\int_D a_1 \nabla u_1 \nabla v - \int_D a_2 \nabla u_2 \nabla v = \int_D fv - \int_D fv = 0, \tag{5.35}$$

ce qui entraîne

$$\int_D a_1 (\nabla u_1 - \nabla u_2) \nabla v = \int_D (a_2 - a_1) \nabla u_2 \nabla v. \tag{5.36}$$

En posant $w = u_1 - u_2$ et en prenant $v = w$, on obtient ainsi

$$r \|w\|_V^2 \leq \int_D a_1 |\nabla w|^2 = \int_D (a_2 - a_1) \nabla u_2 \nabla w \leq \|a_1 - a_2\|_{L^\infty} \|w\|_V \|u_2\|_V. \tag{5.37}$$

On conclut en divisant par $r\|w\|_V$ et en remarquant que $\|u_2\|_V \leq M$. □

Ce résultat nous montre immédiatement que $y \mapsto u(y)$ est Lipschitizienne dans Y en chacune des variables y_j puisque si on pose

$$e_j = (0, \dots, 0, 1, 0, \dots), \tag{5.38}$$

la suite de Kroenecker avec 1 en position j, on a

$$\|a(y + te_j) - a(y)\|_{L^\infty} = |t|\|\psi_j\|_{L^\infty}, \tag{5.39}$$

et par conséquent, en supposant que y et $y + te_j$ sont dans Y, on obtient

$$\|u(y + te_j) - u(y)\|_V \leq C|t|, \quad C = \frac{M}{r}\|\psi_j\|_{L^\infty}. \tag{5.40}$$

Nous allons l'utiliser pour aller plus loin et établir la différentiabilité de $y \mapsto u(y)$.

On peut tout d'abord identifier la dérivée partielle

$$\partial_j u(y) = \frac{\partial u}{\partial y_j}(y), \tag{5.41}$$

par un calcul formel en supposant a-priori que celle-ci est bien définie : on dérive par rapport à y_j la formulation variationelle

$$\int_D a(y)\nabla u(y)\nabla v = \int_D fv. \tag{5.42}$$

Le terme de droite ne dépend pas de y et on obtient ainsi

$$\int_D a(y)\nabla\partial_j u(y)\nabla v + \int_D \partial_j a(y)\nabla u(y)\nabla v = 0, \tag{5.43}$$

soit

$$\int_D a(y)\nabla\partial_j u(y)\nabla v = -\int_D \psi_j \nabla u(y)\nabla v, \quad v \in V. \tag{5.44}$$

En d'autre terme $\partial_j u(y) \in V$ est solution d'un problème aux limites similaire à celui dont $u(y)$ est solution, mais avec une forme linéaire modifiée qui dépend de $u(y)$:

$$A_y(\nabla\partial_j u(y), v) = L_{j,y}(v), \qquad L_{j,y}(v) = -\int_D \psi_j \nabla u(y)\nabla v. \tag{5.45}$$

Pour rendre ce calcul plus rigoureux, appelons $u_j(y) \in V$ la solution de

$$\int_D a(y)\nabla u_j(y)\nabla v = -\int_D \psi_j \nabla u(y)\nabla v, \tag{5.46}$$

et montrons qu'il s'agit bien de la dérivée partielle au sens où

$$u_j(y) = \lim_{t\to 0} u_{t,j}(y), \quad u_{t,j}(y) = \frac{1}{t}\Big(u(y + te_j) - u(y)\Big), \tag{5.47}$$

où la limite s'entend dans V. Pour cela, on écrit

$$\int_D a(y)\nabla u(y)\nabla v = \int_D a(y+te_j)\nabla u(y+te_j)\nabla v, \quad v \in V, \tag{5.48}$$

ce qui par $a(y+te_j) = a(y) + t\psi_j$ nous donne

$$\int_D a(y)\nabla u_{t,j}(y)\nabla v = -\int_D \psi_j \nabla u(y+te_j)\nabla v, \quad v \in V. \tag{5.49}$$

On a ainsi

$$\int_D a(y)\nabla(u_{t,j}(y) - u_j(y))\nabla v = \int_D \psi_j \nabla(u(y) - u(y+te_j))\nabla v, \quad v \in V. \tag{5.50}$$

En prenant $v = u_{t,j}(y) - u_j(y)$ on obtient ainsi

$$\|u_{t,j}(y) - u_j(y)\|_V \leq r^{-1}\|\psi_j\|_{L^\infty}\|u(y) - u(y+te_j)\|_V \to 0 \quad \text{quand} \quad t \to 0. \tag{5.51}$$

Nous avons ainsi identifié les dérivée partielles d'ordre 1. En particulier les coefficients u_ν pour $\nu = e_j$, sont donnés

$$u_{e_j} = \partial_j u(0), \tag{5.52}$$

solution des problèmes aux limites

$$\int_D \overline{a}\nabla u_{e_j}\nabla v = -\int_D \psi_j \nabla u_0 \nabla v, \quad v \in V. \tag{5.53}$$

où $u_0 = u(0)$ est solution de

$$\int_D \overline{a}\nabla u_0 \nabla v = \int_D fv, \quad v \in V. \tag{5.54}$$

On peut itérer ce calcul pour justifier l'existence des dérivées d'ordre supérieur, qu'on calcule formellement en appliquant la dérivée ∂^ν à la formulation variationelle. Comme les dérivées d'ordre supérieur à deux de $a(y)$ sont nulles du faite de la forme affine, on obtient

$$\int_D a(y)\nabla\partial^\nu u(y)\nabla v + \sum_{j\,:\,\nu_j \neq 0} \nu_j \int_D \psi_j \nabla \partial^{\nu - e_j} u(y)\nabla v = 0. \tag{5.55}$$

En divisant par $\nu! = \nu_j(\nu - e_j)!$, et en se placant en $y = 0$, on obtient ainsi la formule

$$\int_D \overline{a}\nabla u_\nu \nabla v = -\sum_{j\,:\,\nu_j \neq 0} \int_D \psi_j \nabla u_{\nu - e_j}\nabla v, \quad v \in V, \tag{5.56}$$

qui nous montre que les coefficients de la série entière peuvent se calculer par récurrence, par la résolution d'une suite de problèmes aux limites.

Il est bien connu que l'existence des dérivées en 0 à tous les ordres ne garantie pas la convergence de la série entière, et que même si cette série converge, la limite peut ne pas être égale à la fonction qu'on a dérivé en 0 : le contre-exemple le plus simple est celui

d'une fonction d'une variable dont toutes les dérivées s'annullent en 0 mais qui n'est pas nulle en dehors de 0. En dimension $d = 1$, un critère suffisant pour avoir l'identité

$$\varphi(t) = \sum_{n\geq 0} \frac{1}{n!}\varphi^{(n)}(0)t^n, \quad t \in [-1, 1], \tag{5.57}$$

est que la fonction φ soit prolongeable en une fonction à variable et valeur complexe holomorphe sur un voisinage ouvert du disque unité $\{|z| \leq 1\}$, c'est à dire que ce prolongement admette une dérivée complexe

$$\varphi'(z) = \lim_{h\to 0} h^{-1}(\varphi(z + h) - \varphi(z)), \tag{5.58}$$

pour $|z| \leq 1 + \varepsilon$, la limite étant au sens de $h \to 0$ dans $\mathbb{C}$. La convergence de la série dans (5.59) est alors ponctuelle mais aussi uniforme et normale, puisque son rayon de convergence est supérieur à 1. Ce résulat est aussi valable pour les fonctions allant de $\mathbb{C}$ dans $\mathbb{C}^N$, où plus généralement de $\mathbb{C}$ dans V un espace de Banach complexe, avec la même définition pour la dérivée complexe.

De même, pour une fonction de plusieurs variables $y = (y_1, \ldots, y_d) \mapsto \varphi(y)$, l'identité

$$\varphi(y) = \sum_{\nu\in\mathbb{N}^d} \frac{1}{\nu!}\partial_\nu\varphi(0)y^\nu, \quad y \in [-1, 1]^d, \tag{5.59}$$

est valable dès que φ admet un prolongement holomorphe sur un voisinage ouvert du polydisque unité

$$\{z = (z_1, \ldots, z_d) \,:\, |z_j| \leq 1\} = \otimes_{j=1}^d \{|z_j| \leq 1\}, \tag{5.60}$$

c'est à dire que ce prolongement admet des dérivées partielles complexes

$$\partial_j\varphi(z) = \lim_{h\to 0} h^{-1}(\varphi(z + he_j) - \varphi(z)), \tag{5.61}$$

pour $|z_j| \leq 1 + \varepsilon$, la limite étant au sens de $h \to 0$ dans $\mathbb{C}$.

Dans notre cas la fonction $u : Y \to V$ dépend d'un nombre infini de variables mais on va voir qu'on peut faire converger la série entière en augmentant progressivement le nombre de variables actives. Il nous faut tout d'abord définir un prolongement

$$z = (z_j)_{j\geq 1} \mapsto u(z), \tag{5.62}$$

pour lequel on a des dérivées partielles complexe en chaque variable. Ceci est rendu possible grâce à une extension du théorème de Lax-Milgram au cas où a est à valeur complexe : le résultat reste valable (**exercice**) en modifiant les hypothèses de continuité et d'ellipticité suivant

$$|A(u, v)| \leq R\|u\|_V\|v\|_V \quad \text{et} \quad \Re(A(u, u)) \geq r\|u\|_V^2, \quad u, v \in V. \tag{5.63}$$

On applique cela à

$$A_z(u, v) = \int_D a(z)\nabla u\overline{\nabla v} \quad \text{et} \quad L(v) = \int_D f\overline{v}, \quad a(z) = \overline{a} + \sum_{j\geq 1} z_j\ \psi_j. \tag{5.64}$$

On remarque que l'hypothèse d'ellipticité uniforme (5.16) peut se réexprimer de manière équivalente (**exercice**) par

$$\sum_{j\geq 1} |\psi_j(x)| \leq \overline{a}(x) - r \quad \text{et} \quad \sum_{j\geq 1} |\psi_j(x)| \leq R - \overline{a}(x), \quad x \in D. \tag{5.65}$$

Ceci entraine immédiatement que la fonction $a(z)$ vérifie

$$|a(z)| \leq R \quad \text{et} \quad \Re(a(z)) \geq r, \tag{5.66}$$

si on a $|z_j| \leq 1$ pour tout $j \geq 1$. Ces inégalités permettent de vérifier les hypothèses (5.63) pour A_z, et de définir ainsi une extension

$$z \mapsto u(z), \tag{5.67}$$

de l'application solution sur le polydisque

$$\mathcal{Y} = \otimes_{j\geq 1}\{|z_j| \leq 1\}. \tag{5.68}$$

L'existence des dérivées partielles complexes $\partial_j u(z)$ se démontre comme dans le cas des variables réelles : celles-ci sont solutions des problèmes aux limites

$$\int_D a(z)\nabla\partial_j u(z)\overline{\nabla v} = -\int_D \psi_j \nabla u(z)\overline{\nabla v}, \quad v \in V. \tag{5.69}$$

Nous sommes à présent en position de prouver la convergence de la série entière (5.31) vers $u(y)$. Pour cela, on effectue une troncature des variables en introduisant pour $J \geq 1$ la fonction de J variables

$$u_J(y_1, \dots, y_J) = u(y_1, \dots, y_J, 0, 0, \dots), \tag{5.70}$$

définie sur $[-1, 1]^J$ et qui admet une extension holomorphe sur $\otimes_{j=1}^d\{|z_j| \leq 1+\varepsilon\}$ pour $\varepsilon > 0$ suffisament petit (**exercice**). On a donc un développement en série entière uniformément convergent

$$u_J(y_1, \dots, y_J) = \sum_{\nu\in\mathbb{N}^J} \frac{1}{\nu!}\partial^\nu u_J(0, \dots, 0)\prod_{j=1}^J y_j^{\nu_j}. \tag{5.71}$$

On peut écrire $\partial^\nu u_J(0, \dots, 0) = \partial^{\overline{\nu}} u(0)$ en posant $\overline{\nu} = (\nu_1, \dots, \nu_J, 0, 0, \dots)$ pour $\nu = (\nu_1, \dots, \nu_J)$. Ceci nous montre que pour tout $y = (y_1, \dots, y_J, y_{J+1}, \dots) \in Y$, on a

$$u_J(y_1, \dots, y_J) = \sum_{\nu\in\mathcal{F}_J} u_\nu y^\nu, \tag{5.72}$$

où $\mathcal{F}_J$ désigne l'ensemble des multi-indices à support dans $\{1, \dots, J\}$,

$$\mathcal{F}_J = \{\nu \in \mathcal{F} \ : \ \nu_j = 0, \ j > J\}. \tag{5.73}$$

Le lemme de stabilité (5.1) nous indique que

$$\|u(y) - u_J(y)\|_V \leq C\|a(y) - a_J(y)\|_{L^\infty}, \tag{5.74}$$

et d'après l'hypothèse de convergence uniforme (5.4), nous avons donc

$$u(y) = \lim_{J\to+\infty} u_J(y) = \lim_{J\to\infty} \sum_{\nu\in\mathcal{F}_J} u_\nu y^\nu = \sum_{\nu\in\mathcal{F}} u_\nu y^\nu, \tag{5.75}$$

et cette convergence est uniforme sur Y.

5.3 Un résultat de parcimonie

Nous venons de prouver la convergence de la série entière au sens d'un procédé de sommabilité spécifique : on somme sur $\mathcal{F}_J$ et on fait tendre $J \to \infty$. Nous allons voir que, sous des hypothèses supplémentaires, la convergence est en fait inconditionnelle au sens où on a

$$\sum_{\nu \in \mathcal{F}} \|u_\nu\|_V < \infty, \tag{5.76}$$

ce qui nous permet de restreindre la série à n'importe quel ensemble $\Lambda \subset \mathcal{F}$ et d'écrire

$$\sup_{y \in Y} \left\| u(y) - \sum_{u \in \Lambda} u_\nu y^\nu \right\|_V = \sup_{y \in Y} \left\| \sum_{u \notin \Lambda} u_\nu y^\nu \right\|_V \leq \sum_{u \notin \Lambda} \|u_\nu\|_V. \tag{5.77}$$

Cette estimation est assez brutale puisqu'on a utilisé l'inégalité triangulaire, mais elle nous permet de ramener le problème du choix d'un ensemble Λ_n de cardinal n pour définir le polynôme u_n, au choix de celui qui minimize l'erreur de meilleure approximation à n-termes de la suite $(\|u_\nu\|_V)_{\nu \in \mathcal{F}}$ en norme $\ell^1(\mathcal{F})$. Le choix optimal consiste à prendre l'ensemble Λ_n des multi-indices correspondant aux n plus grands $\|u_\nu\|_V$. Avec ce choix, la théorie de l'approximation à n-termes nous donne aussi une indication sur la vitesse de convergence grâce au Theorème 2.6 appliqué avec $q = 1$ et $p < 1$:

$$(\|u_\nu\|_V)_{\nu \in \mathcal{F}} \in \ell^p(\mathcal{F}) \quad \Longrightarrow \quad \sup_{y \in Y} \|u(y) - u_n(y)\|_V \leq \sum_{u \notin \Lambda_n} \|u_\nu\|_V \leq C n^{-s}, \quad s = \frac{1}{p} - 1, \tag{5.78}$$

avec $C = (\sum_{\nu \in \mathcal{F}} \|u_\nu\|_V^p)^{1/p}$. Nous sommes ainsi ramené à étudier le problème suivant :

Pour quelles valeurs de $p > 0$ la suite $(\|u_\nu\|_V)_{\nu \in \mathcal{F}}$ est ℓ^p-sommable ?

Le résultat suivant, établi en 2011 par Cohen-DeVore-Schwab, apporte une réponse générale sous la forme d'une condition sur les fonctions $(\psi_j)_{j \geq 1}$.

Théorème 5.1 *Sous la condition d'ellipticité uniforme (5.16), on a pour tout $p < 1$ l'implication*

$$\sum_{j \geq 1} \|\psi_j\|_{L^\infty}^p < \infty \quad \Longrightarrow \quad \sum_{\nu \in \mathcal{F}} \|u_\nu\|_V^p < \infty. \tag{5.79}$$

Une première intuition de ce résultat vient de la sous-suite des termes d'ordres $|\nu| = 1$ qui correspondent à $\nu = e_j$, pour laquelle l'estimation (5.51) nous montre que

$$\|u_{e_j}\|_V \leq \frac{M}{r} \|\psi_j\|_{L^\infty}, \tag{5.80}$$

ce qui rend immédiat l'implication si on se restreint uniquement à ces termes. La difficulté est ici qu'on cherche à contrôler la norme ℓ^p pour la totalité des $\|u_\nu\|_V$. Nous allons donner la preuve dans un cas simple (lorsque les ψ_j sont à supports disjoints), la preuve générale pouvant être consultée dans la référence [3] citée en introduction. Avant d'aborder la preuve, discutons le sens et les implications de ce résultat.

Une première conséquence est que sous la condition $(\|\psi_j\|_{L^\infty})_{j\geq 1} \in \ell^p(\mathbb{N})$, pour tout $n \geq 1$ il existe un ensemble de multi-indice $\Lambda_n \subset \mathcal{F}$ de cardinal n tel que

$$\|u - u_n\|_{L^\infty(Y,V)} = \sup_{y\in Y} \|u(y) - u_n(y)\|_V \leq Cn^{-s}, \quad s = \frac{1}{p} - 1 > 0, \tag{5.81}$$

où on a posé

$$u_n(y) = \sum_{\nu\in\Lambda_n} u_\nu y^\nu \quad \text{et} \quad C = \Big(\sum_{\nu\in\mathcal{F}} \|u_\nu\|_V^p\Big)^{1/p}. \tag{5.82}$$

On a ainsi controurné la malédiction des grandes dimension puisqu'on obtient une vitesse d'approximation polynomiale $\mathcal{O}(n^{-s})$ bien que le nombre de variables soit infini. Ceci est rendu possible à la fois par le fait que l'application $y \mapsto u(y)$ est régulière, mais aussi que sa dépendance en les différentes variables est fortement anisotrope, comme le montre par exemple les estimations (5.80) sur les dérivées partielles d'ordre 1 : quand $\|\psi_j\|_{L^\infty}$ devient petit, les variations suivant y_j deviennent faibles. L'approximation à n termes joue aussi un rôle essentiel pour obtenir la vitesse n^{-s} qui ne serait généralement pas obtenue avec d'autres choix d'ensembles Λ_n.

Voici une autre conséquence de ce résultat : en définissant le sous-espace de V de dimension n

$$V_n := \text{vect}\{u_\nu \ : \ \nu \in \Lambda_n\}, \tag{5.83}$$

on a $u_n(y) \in V_n$ pour tout $y \in Y$ ce qui montre que

$$\sup_{y\in Y} \min_{v\in V_n} \|u(y) - v\|_V \ \leq Cn^{-s}. \tag{5.84}$$

De manière équivalente, si $\mathcal{K} = \{u(y) \ : \ y \in Y\}$ est la variété solution, on a

$$\text{dist}(\mathcal{K}, V_n) = \sup_{u\in\mathcal{K}} \min_{v\in V_n} \|u - v\|_V \leq Cn^{-s}. \tag{5.85}$$

Le résultat d'approximation polynomiale nous permet donc d'affirmer que l'épaisseur de Kolmogorov de $\mathcal{K}$ vérifie

$$d_n(\mathcal{K})_V \leq Cn^{-s}. \tag{5.86}$$

Notons que c'est uniquement une borne supérieure, il est possible que d_n décroisse encore plus rapidement que n^{-s}.

Afin de prouver le Théorème 5.1, il faut établir des estimations des coefficients $\|u_\nu\|_V$. Une approche possible se fonde sur l'holomorphie du prolongement $z \mapsto u(z)$ qu'on a définit précèdemment. On remarque d'abord que ce prolongement holomorphe peut être défini au dela du polydisque $\mathcal{Y}$: si $\rho = (\rho_j)_{j\geq 1}$ est une suite de nombres supérieurs à 1, on lui associe le polydisque

$$\mathcal{Y}_\rho := \{z = (z_j)_{j\geq 1} \ : \ |z_j| \leq \rho_j\} = \otimes_{j\geq 1}\{|z_j| \leq \rho_j\}. \tag{5.87}$$

Pour $\delta > 0$, si une suite positive $(\rho_j)_{j\geq 1}$ est telle que

$$\sum_{j\geq 1} \rho_j |\psi_j(x)| \leq \overline{a}(x) - \delta, \quad x \in D, \tag{5.88}$$

on voit que

$$z \in \mathcal{Y}_\rho \implies \Re(a(z)) \geq \delta \quad \text{et} \quad |a(z)| \leq \overline{R} = 2R - \delta, \tag{5.89}$$

ce qui nous assure l'existence de la solution $u(z)$ et l'holomorphie de $z \mapsto u(z)$ sur $\mathcal{Y}_\rho$. On voit en particulier que si l'on prend $0 < \delta < r$, on aura la possibilité de choisir des nombres ρ_j supérieurs à 1 : puisqu'on sait déjà que $\sum_{j\geq 1} |\psi_j(x)| \leq \overline{a}(x) - r$ il suffit d'avoir

$$\sum_{j\geq 1} (\rho_j - 1)|\psi_j(x)| \leq r - \delta, \qquad x \in D. \tag{5.90}$$

Dans toute la suite on prendra $\delta = \frac{r}{2}$ et on dira qu'une suite ρ de nombres supérieurs ou égaux à 1 est *admissible* si elle vérifie la propriété (5.88) pour cette valeur de δ. En désignant par $\mathcal{A}$ l'ensemble des suites admissibles, on a donc l'estimation a-priori

$$\|u(z)\|_V \leq 2M = \frac{C_P \|f_{L^2}}{\delta}, \quad z \in \mathcal{Y}_\rho, \quad \rho \in \mathcal{A}. \tag{5.91}$$

On rappelle ensuite la formule de Cauchy : si φ est une fonction holomorphe d'une variable définie sur un ouvert $\Omega \subset \mathbb{C}$, à valeur dans $\mathbb{C}$ ou plus généralement dans un espace de Banach complexe V, et si Γ est une courbe fermée régulière contenue dans Ω et $z \in \mathbb{C}$ est à l'intérieur du domaine circonscrit par Γ, on a

$$\varphi(z) = \frac{1}{2i\pi} \int_\Gamma \frac{\varphi(\tilde{z})}{\tilde{z} - z} d\tilde{z}. \tag{5.92}$$

Cette formule permet en particulier de calculer les dérivées de φ à tous les ordres en dérivant $(\tilde{z} - z)^{-1}$ dans l'intégrale, ce qui donne

$$\varphi^{(m)}(z) = \frac{1}{2i\pi} m! \int_\Gamma \frac{\varphi(\tilde{z})}{(\tilde{z} - z)^{m+1}} d\tilde{z}. \tag{5.93}$$

Dans le cas où φ est holomorphe sur un voisinage ouvert d'un disque $D_\lambda = \{|z| \leq \lambda\}$, on obtient ainsi en $z = 0$,

$$|\varphi^{(m)}(0)| \leq \frac{1}{2\pi} m! \int_{|\tilde{z}|=\lambda} \frac{|\varphi(z')|}{|\tilde{z}|^{m+1}} dz' \leq C m! \lambda^{-m}, \quad C = \max_{z \in D_\lambda} |\varphi(z)|. \tag{5.94}$$

Dans le cas d'une fonction à valeur dans un espace de Banach complexe V, on obtient de la même manière

$$\|\varphi^{(m)}(0)\|_V \leq C m! \lambda^{-m}, \quad C = \max_{z \in D_\lambda} \|\varphi(z)\|_V. \tag{5.95}$$

La formule de Cauchy se généralise par récurrence (**exercice**) à une fonction holomorphe de plusieurs variables complexe $\varphi(z_1, \ldots, z_d)$ lorsque $\Omega = \Omega_1 \times \cdots \times \Omega_d$ et $\Gamma = \Gamma_1 \times \cdots \times \Gamma_d$ en écrivant

$$\varphi(z) = \frac{1}{(2i\pi)^d} \int_{\Gamma_1} \int_{\Gamma_2} \cdots \int_{\Gamma_d} \frac{\varphi(\tilde{z}_1, \ldots, \tilde{z}_d)}{(\tilde{z}_1 - z_1) \cdots (\tilde{z}_d - z_d)} d\tilde{z}_1 \ldots d\tilde{z}_d. \tag{5.96}$$

On peut effectuer le calcul de la dérivée partielle en appliquant le même procédé en chacune des variables. Si φ est holomorphe sur un voisinage ouvert du polydisque $D_\lambda = D_{\lambda_1} \times \cdots \times D_{\lambda_d}$, pour $\lambda = (\lambda_1, \ldots, \lambda_d)$, on obtient ainsi pour tout $\nu \in \mathbb{N}^d$ l'estimation

$$\|\partial^\nu \varphi(0)\|_V \leq C\nu!\lambda^{-\nu}, \quad C = \max_{z \in D_\lambda} \|\varphi(z)\|_V, \tag{5.97}$$

où $\lambda^{-\nu} = \prod_{j=1}^d \lambda_j^{-\nu_j}$. On peut appliquer cela à $z \mapsto u(z)$, en remarquant que si $\nu \in \mathcal{F}$ est à support dans $\{1, \ldots, J\}$, on peut identifier $\partial^\nu u(0)$ à la dérivée partielle $\partial^\nu u_J(0)$, où u_J est la fonction de J variables $u_J(z_1, \ldots, z_J) = u(z_1, \ldots, z_J, 0, 0, \ldots)$. On obtient ainsi

$$\|\partial^\nu u(0)\|_V \leq C\nu!\rho^{-\nu}, \quad C = \max_{z \in \mathcal{Y}_\rho} \|u(z)\|_V, \tag{5.98}$$

pour toute suite ρ telle que $z \mapsto u(z)$ est bornée sur $\mathcal{Y}_\rho$ et holomorphe en chaque variable sur un voisinage ouvert de $\mathcal{Y}_\rho$. Dans les cas des suites admissibles on obtient ainsi

$$\rho \in \mathcal{A} \implies \|u_\nu\|_V \leq 2M\rho^{-\nu}, \tag{5.99}$$

avec $\rho^{-\nu} = \prod_{j \geq 1} \rho_j^{-\nu_j}$. Comme on peut utiliser n'importe quelle suite admissible, on a obtenu l'estimation

$$\|u_\nu\|_V \leq e_\nu := 2M \inf\{\rho^{-\nu} \ : \ \rho \in \mathcal{A}\}. \tag{5.100}$$

Le calcul de l'estimateur e_ν fait apparaitre pour chaque $\nu \in \mathcal{F}$ un problème d'optimisation :

$$\text{minimiser} \quad \prod_{j \geq 1} \rho_j^{-\nu_j} \quad \text{sous les contraintes} \quad \sum_{j \geq 1} \rho_j |\psi_j(x)| \leq \overline{a}(x) - \frac{r}{2}, \quad x \in D. \tag{5.101}$$

La résolution exacte de ce problème de minimisation est difficile à cause de la contraintes qui dépendent du point x et couplent l'ensemble des ρ_j. Il est ainsi difficile de caractériser la suite $\rho = \rho(\nu)$ qui réalise le minimum et de trouver ainsi la valeur exacte de e_ν.

Le calcul exact est cependant possible dans le cas particulier où les ψ_j sont à support disjoints, par exemple si ils sont de la forme $\psi_j = \alpha_j \chi_{D_j}$ avec $\alpha_j \in \mathbb{R}_+$ et $(D_j)_{j \geq 1}$ une partition de D. En effet dans ce cas les contraintes se séparent deviennent

$$\rho_j |\psi_j(x)| \leq \overline{a}(x) - \frac{r}{2}, \quad x \in D, \quad j \geq 1. \tag{5.102}$$

On minimise donc $\rho^{-\nu}$ en prenant pour chaque j la valeur maximale de ρ_j vérifiant cette contrainte, c'est à dire

$$\rho_j = \min_{x \in D} \frac{\overline{a}(x) - \frac{r}{2}}{|\psi_j(x)|}, \tag{5.103}$$

et on note que cette valeur est supérieure à 1 puisque $|\psi_j(x)| \leq \overline{a}(x) - r$ d'après l'hypothèse d'ellipticité uniforme. Notons aussi que dans ce cas la suite optimale $\rho = (\rho_j)_{j \geq 1}$ est indépendante du multi-indice ν.

Nous allons utiliser cette estimation pour démontrer le Théorème 5.1 dans ce cas. On introduit la suite $b = (b_j)_{j \geq 1}$ inverse de ρ c'est à dire

$$b_j = \rho_j^{-1} = \max_{x \in D} \frac{|\psi_j(x)|}{\overline{a}(x) - \frac{r}{2}}, \tag{5.104}$$

qui vérifie $0 < b_j < 1$ pour tout j et d'autre part

$$b_j \leq C\|\psi_j\|_{L^\infty}, \quad C = \frac{2}{r}, \tag{5.105}$$

puisque l'hypothèse d'ellipticité uniforme entraîne en particulier $\overline{a}(x) \geq r$. Par conséquent l'hypothèse $(\|\psi_j\|_{L^\infty})_{j\geq 1} \in \ell^p(\mathbb{N})$ entraine que $b \in \ell^p(\mathbb{N})$. En particulier $b_j \to 0$ ce qui nous montre aussi que

$$\|b\|_{\ell^\infty} = \max_{j\geq 1} b_j < 1. \tag{5.106}$$

L'estimation sur les $\|u_\nu\|_V$ pouvant s'écrire

$$\|u_\nu\|_V \leq 2Mb^\nu, \tag{5.107}$$

le théorème découle du résultat général suivant.

Lemme 5.2 *Soit $b = (b_j)_{j\geq 1}$ une suite de nombre positifs. On a pour tout $0 < p < \infty$ alors*

$$\|b\|_{\ell^\infty} < 1 \quad \text{et} \quad b \in \ell^p(\mathbb{N}) \quad \Longleftrightarrow \quad (b^\nu)_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F}). \tag{5.108}$$

.

Preuve : Si $\|b\|_{\ell^\infty} < 1$, on a pour chaque j l'égalité

$$\sum_{n\geq 0} b_j^{pn} = \frac{1}{1-b_j^p}. \tag{5.109}$$

D'autre part, on peut écrire par développement limité en $x = 0$ de $\ln(1-x)$,

$$\ln\Big(\frac{1}{1-b_j^p}\Big) = -\ln(1-b_j^p) \leq Cb_j^p, \quad C = \frac{1}{1-\|b\|_{\ell^\infty}^p} < \infty. \tag{5.110}$$

Ceci nous montre que le produit

$$\prod_{j\geq 1}\Big(\sum_{n\geq 0} b_j^{pn}\Big) = \prod_{j\geq 1} \frac{1}{1-b_j^p}, \tag{5.111}$$

est convergent, et borné par $\exp(C\|b\|_{\ell^p}^p)$. Or en développant le produit à gauche tronqué à l'ordre J, on trouve exactement

$$\prod_{j=1}^{J}(\sum_{n\geq 0} b_j^{pn}) = \sum_{n_1,\ldots,n_J\in\mathbb{N}} \prod_{j=1}^{J} b_j^{pn_j} = \sum_{\nu\in\mathcal{F}_J} (b^\nu)^p. \tag{5.112}$$

En faisant tendre J vers $+\infty$, on obtient ainsi

$$\sum_{\nu\in\mathcal{F}} (b^\nu)^p = \prod_{j\geq 1} \frac{1}{1-b_j^p} \leq \exp(C\|b\|_{\ell^p}^p) < \infty. \tag{5.113}$$

ce qui nous montre l'implication dans le premier sens, qui est celle dont on a besoin pour démontrer le théorème. L'implication inverse est immédiate, en remarquant que la suite

$(b^\nu)_{\nu\in\mathcal{F}}$ contient en particulier tous les b_j^n pour $j \geq 1$ et $n \geq 0$. $\square$

La preuve du Theorème 5.1 dans le cas général est aussi fondée sur l'estimation de $\|u_\nu\|_V$ par e_ν, mais nécessite pour chaque ν une suite $\rho(\nu) \in \mathcal{A}$ astucieusement choisie de manière à pouvoir établir que la suite $(\rho(\nu)^{-\nu})_{\nu\in\mathcal{F}}$ appartient à $\ell^p(\mathcal{F})$.

Par une méthode qui n'utilise pas l'analyse complexe, il aussi possible d'améliorer l'estimation des $\|u_\nu\|_V$ en prouvant le résultat suivant : pour toute suite $\rho \in \mathcal{A}$, on a

$$\sum_{\nu\in\mathcal{F}} \Big(\rho^\nu \|u_\nu\|_V\Big)^2 \leq C < \infty, \tag{5.114}$$

où la constante C dépend de (r, R, M). Ce résultat se prouve en utilisant la norme $\|v\|^2 = \int \overline{a}|\nabla v|^2$ qui est équivalente à $\|v\|_V^2$, et en utilisant la formule de récurrence (5.56) sur les u_ν pour établir (**exercice difficile**) la propriété de contraction

$$\sum_{|\nu|=k+1} \Big(\rho^\nu \|u_\nu\|\Big)^2 \leq \kappa \sum_{|\nu|=k} \Big(\rho^\nu \|u_\nu\|\Big)^2, \tag{5.115}$$

où $\kappa < 1$ depend de (r, R, M). L'estimation (5.114) entraîne clairement celle des $\|u_\nu\|_V$ par $\rho^{-\nu}$ mais elle est un peu plus forte. Pour $p < 2$, on peut en particulier utiliser l'inégalité de Hölder pour écrire

$$\Big(\sum_{\nu\in\mathcal{F}} \|u_\nu\|_V^p\Big)^{1/p} \leq \Big(\sum_{\nu\in\mathcal{F}} \Big(\rho^\nu \|u_\nu\|_V\Big)^2\Big)^{1/2} \Big(\sum_{\nu\in\mathcal{F}} \rho^{-q\nu}\Big)^{1/q}, \tag{5.116}$$

avec $\frac{1}{q} = \frac{1}{p} - \frac{1}{2}$. En combinant ceci avec le Lemme 5.2, on obtient le résultat suivant.

Théorème 5.2 *Sous la condition d'ellipticité uniforme (5.16), si il existe $\rho \in \mathcal{A}$ telle que la suite inverse b appartient à $\ell^q(\mathbb{N})$ pour un $q < \infty$, alors $(\|u_\nu\|_V)_{\nu\in\mathcal{F}}$ appartient à $\ell^p(\mathcal{F})$ avec $\frac{1}{p} = \frac{1}{q} + \frac{1}{2}$.*

Dans le cas des ψ_j à supports disjoints, ce résultat apporte immédiatement une amélioration sur le Théorème 5.1 puisqu'on a vu qu'on peut prendre une suite $\rho \in \mathcal{A}$ telle que $b_j \leq C\|\psi_j\|_{L^\infty}$. On obtient ainsi dans ce cas l'implication

$$\sum_{j\geq 1} \|\psi_j\|_{L^\infty}^q < \infty \quad \Longrightarrow \quad \sum_{\nu\in\mathcal{F}} \|u_\nu\|_V^p < \infty, \quad \frac{1}{p} = \frac{1}{q} + \frac{1}{2}. \tag{5.117}$$

On ne peut pas obtenir la même amélioration lorsque les ψ_j ont des supports qui se recouvrent de manière arbitraire.

Il existe de multiples variantes et généralisations des Théorèmes 5.1 et 5.2, en particulier à d'autres modèles d'EDP que celui fourni par l'équation elliptique (5.1) et la dépendance affine (5.2) en les paramètres y_j. Signalons cependant que des difficultés sérieuses se présentent dans le cas des équations hyperboliques paramétriques (transport linéaire ou non-linéaire, ondes...) pour lesquels les développements polynomiaux ne sont pas aussi efficaces, tout comme les méthodes de bases réduites.

On peut aussi établir des résultats similaires pour d'autres types d'expansions polynomiales que les séries de Taylor, en particulier pour les décompositions en polynômes orthogonaux. Si on considère en particulier la base des polynômes de Legendre univariés $(L_n)_{n\geq 0}$ que l'on normalise dans $L^2([-1,1],\frac{dt}{2})$, on définit par tensorisation une famille

$$L_\nu(y) = \prod_{j\geq 1} L_{\nu_j}(y_j), \qquad \nu \in \mathcal{F}, \tag{5.118}$$

qui est une base orthonormale de l'espace $L^2(Y,d\mu)$ où $d\mu$ est la mesure produit

$$d\mu = \prod_{j\geq 1} \frac{dy_j}{2}, \tag{5.119}$$

c'est à dire la mesure uniforme sur Y. On passe ici sous silence certains détails techniques pour la définition rigoureuse de ces espaces qui sont nécessaire à cause de la dimension infinie de Y et de la mesure produit. L'application $y \mapsto u(y)$ peut être décomposée dans cette base suivant

$$u(y) = \sum_{\nu\in\mathcal{F}} v_\nu L_\nu(y), \qquad v_\nu = \int_Y u(y) L_\nu(y) d\mu, \tag{5.120}$$

où la série converge dans l'espace $L^2(Y,V,d\mu)$ qui est muni de la norme

$$\|u\|^2_{L^2(Y,V,d\mu)} = \int_Y \|u(y)\|^2_V d\mu = \sum_{\nu\in\mathcal{F}} \|v_\nu\|^2_V. \tag{5.121}$$

On peut montrer que sous les mêmes hypothèses que celles des Théorèmes 5.1 et 5.2, la suite des coefficients de Legendre $(\|v_\nu\|_{\nu\in\mathcal{F}})$ appartient à $\ell^p(\mathcal{F})$. Si on définit cette fois le polynôme d'approximation par

$$u_n(y) = \sum_{\nu\in\Lambda_n} v_\nu L_\nu(y), \tag{5.122}$$

où Λ_n est l'ensemble des indices correspondant aux n plus grand $\|v_\nu\|_V$, son l'erreur d'approximation vérifie

$$\|u-u_n\|_{L^2(Y,V,d\mu)} = \Big(\sum_{\nu\notin\Lambda_n} \|v_\nu\|^2\Big)^{1/2}. \tag{5.123}$$

En particulier le Theorème 2.6 appliqué avec $q=2$ et $p<2$, nous montre que si on a prouvé que $(\|v_\nu\|_{\nu\in\mathcal{F}}) \in \ell^p(\mathcal{F})$, ceci entraîne l'estimation d'erreur

$$\|u-u_n\|_{L^2(Y,V,d\mu)} \leq Cn^{-s}, \qquad s := \frac{1}{p} - \frac{1}{2}. \tag{5.124}$$

Notons que pour l'erreur en norme $L^\infty(Y,V)$ avec les séries de Taylor, on avait obtenu la valeur $s = \frac{1}{p} - 1$ qui était un peu moins bonne que $\frac{1}{p} - \frac{1}{2}$. Cette amélioration tient au

fait qu'on a fait apparaitre la norme ℓ^2 grâce à l'orthogonalité et l'égalité de Parseval, plutôt que la norme ℓ^1 qu'on avait obtenue par application de l'inégalité triangulaire.

L'estimation d'erreur en norme $L^2(Y, V, d\mu)$ a une interprétation en termes probabilistes lorsque le vecteur y est aléatoire et de loi uniforme sur Y, c'est à dire lorsque les paramètres y_j sont indépendants et uniformément distribués sur $[-1, 1]$: on a alors

$$\mathbb{E}(\|u - u_n\|_V^2) = \|u - u_n\|_{L^2(Y,V,d\mu)}^2 \leq Cn^{-2s}. \tag{5.125}$$

ce qui entraîne en particulier que la quantité

$$\sigma_n^2 := \min_{\dim(V_n)=n} \mathbb{E}(\|u - P_{V_n} u\|_V^2), \tag{5.126}$$

décroît au moins à la vitesse n^{-2s}.